3D Human Action Recognition In Multiple View Scenarios

C.Canton-Ferrer[1], Student Member, IEEE, J.R.Casas[1], Member, IEEE, M.Pardàs[1], Member, IEEE,
M.E.Sargin[2], A.M.Tekalp[2], Fellow, IEEE

[1]Image Processing Group, Technical University of Catalonia, Barcelona, Spain
[2]Multimedia, Vision and Graphics Laboratory, Koç University, Istanbul, Turkey

***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. 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.**

***Index Terms*—Human Gesture Recognition, Information Fusion, 3D Processing, Motion Analysis**

I. Introduction

Analysis of human motion and gesture in image sequences is a topic that has been studied extensively [1] and 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-based analysis from multiple views. Multiple camera systems have been widely 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 [2], [3] , 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 or 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 [4], [5]. Considerably less work has been published on recognizing human gestures using multiple cameras. Monocular 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 or segmentation errors and provide 3D position information as well.

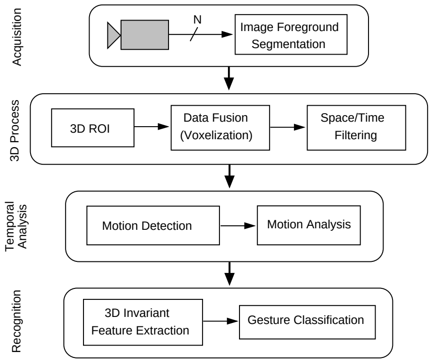


Fig. 1. System flowchart: acquisition, 3D data generation and filtering, motion analysis, robust feature extraction and classification.

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 over these data is performed [4], [6]. This paper explores the complementary approach, first performing a fusion of the incoming data and then extracting 3D motion description features to perform classification.

We propose 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 [4] are presented: Motion History Volume and Motion Energy Volume. Finally, a set of robust 3D invariant statistical moments [7] are computed as a feature vector for classification in a Bayesian framework. Quantitative results for the proposed algorithm are provided as well as a comparison with other motion-based gesture recognition systems [6].

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 disabled people interfaces, body and gait analysis or domotics.

II. System Overview

According to the flowchart depicted in Fig.I the system comprises four data processing modules: image acquisition, 3D data processing and temporal 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 background learning and substraction technique [8], [9]. 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.

A. 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.

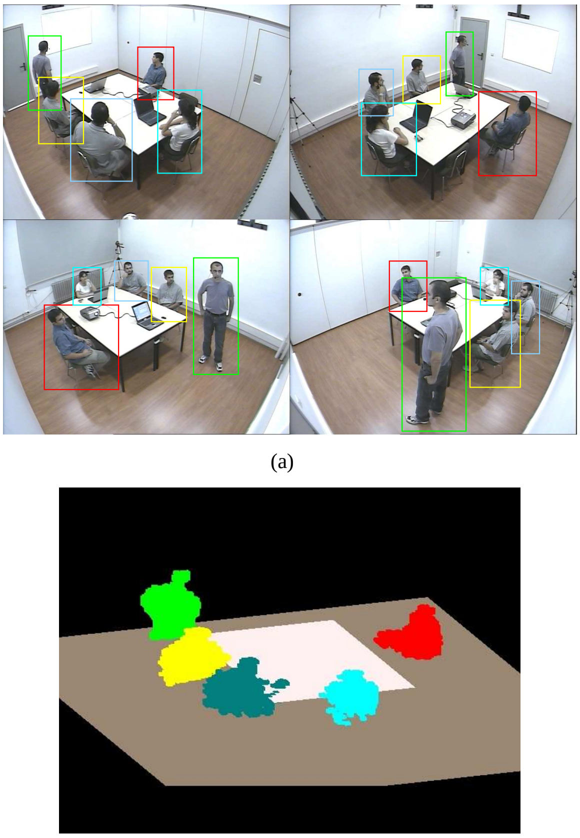
Once foreground regions are extracted from the set of N original images at time t, a set of M 3D points $\mathbf{x}^k$, $0 \leq k < M$, corresponding to the top of each 3D detected volume in the room is obtained by applying a robust Bayesian correspondence algorithm and tracking, as described in [10]. The information given by the established correspondences allows defining a region of interest (ROI) described by a bounding box 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 thus increasing the performance of the whole system.

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

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

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ state for 3D and 2D coordinates respectively, $I_n\left(\tilde{\mathbf{x}},t\right)$ is the segmented image captured by j-th camera, $B^k\left(\mathbf{x},t\right)$ 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 this problem [11] generate new synthetic views by placing virtual cameras in an orthogonal coordinate system related with the center of the action as a data fusion process. 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 N-view silhouette consistency check [12]. This process generates a discrete occupancy representation of the


(a)

(b)

Fig. 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 $\Omega\left(\mathbf{x},t\right)$ proposed in this paper.

3D space (voxels). Information derived from the multiple ROIs allow labeling the voxels as belonging to one person or another. In spite of this fairly simple election of $\mathcal{R}(\cdot)$ compared with more complex reconstruction procedures [13], data fusion still achieves enough accuracy for our purposes.

The data obtained with this 3D reconstruction is corrupted by spurious voxels introduced due to wrong segmentation, camera calibration inaccuracies, etc. Temporal analysis module placed next in the processing chain highly depends on the reliability of the data fusion hence, noise 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 neighbours in both space and time. An example of the output of the whole 3D processing module is depicted in Fig.2(b)

B. *Temporal Analysis Module*

In order to achieve a simple and efficient low level view-dependent motion representation, [4] introduced the concept of Motion History Image (MHV) and Motion Energy Image (MEI). In this paper, we extended the same formulation to represent view-independent 3D motion. Analogously to [4], [5], the binary Motion Energy Volume (MEV) $E_\tau(\mathbf{x},t)$ is defined as:

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

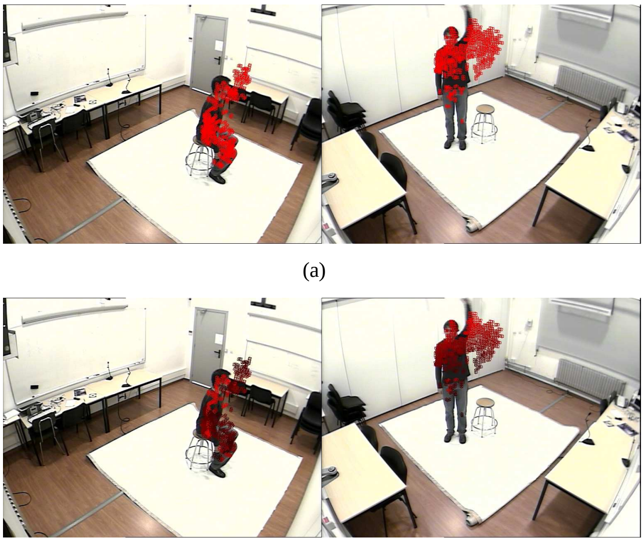
where $\Omega^D(\mathbf{x},t)$ is the binary data set indicating regions of motion. This measure captures the 3D locations where there is motion in the last τ frames. Motion detection captured in $\Omega^D(\mathbf{x},t)$ can be coarsely estimated by a simple forward differentiation among voxel frames. 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.

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,

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

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).


(a)

(b)

Fig. 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.

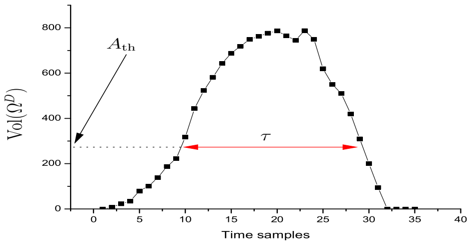


Fig. 4. Estimation of time decay parameter τ of hand waving action by looking at the volume of the motion detection set $\Omega^D(\mathbf{x},t)$.

III. Feature Extraction and Gesture 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 [14] 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 [4], [5], [6]. 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 [15] and can be derived analytically. The reader is referred to Lo and Don's method [7] 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.

Given the computed moment-based features obtained for each of the actions to classify ω_j, $0 \leq j \leq K$, we define a full 10-dimensional feature vector as $\Gamma = [\psi_{\text{MEV}} \;\; \psi_{\text{MHV}}]$. Even though the dimensionality of Γ is very reduced, empty-space related problems arise when estimating class distributions [16]. 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) [16]. By analyzing the training data we noticed that 90% of the variance of the data was achieved by using a dimension reduction to $d = 7$.

The classification method is based on a Bayesian classification criterium assuming that $p(\Gamma|\omega_j)$ is normally distributed. Since the noise in our data is the result of the sum of contribution from a large number of independent sources, Central Limit Theorem grants consistency to the Gaussianity assumption of our data. Indeed, further empiric tests [16] corroborate this assumption. Given an observation represented by Γ, its classification is expressed by the maximum likelihood principle:

$$\arg\max_{\omega_j} p(\omega_j|\Gamma), \tag{4}$$

where the posterior probability of a certain class ω_j given an observation Γ is formally

$$p\left(\omega_j|\Gamma\right) = \frac{p\left(\Gamma|\omega_j\right)\; p\left(\omega_j\right)}{p\left(\Gamma\right)}. \tag{5}$$

Since $p(\omega_j)$ and $p(\Gamma)$ factors are wide and uninformative, Eq.5 can be expressed as

$$p\left(\omega_j|\Gamma\right) \propto p\left(\Gamma|\omega_j\right), \tag{6}$$

where $p\left(\Gamma|\omega_j\right)$ is modeled as a multivariate Gaussian distribution defined by its mean μ and covariance matrix $\mathbf{\Sigma}$. Training data is used to estimate $(\mu, \mathbf{\Sigma})_\mathbf{j}$ for each class in order to compute the class-likelihood discriminant in Eq.4.

IV. 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 wide angle lense 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 the moment invariant based features, actions were recorded in different positions inside the room and facing various orientations.

Quantitative results showed in Table I prove the efficiency of the proposed algorithm to recognize human gestures from the given dataset. In average, we got a $p(\text{error}) = 0.0375$. However, our method is conditioned by the initial foreground segmentation step thus being sensitive to the colours of the clothes of the people in the scene.

TABLE I

Confusion matrix indicating the $p(\text{error})$ of the Bayesian classifier among gesture classes.

	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	0.0	0.0	0.0	0.0	**0.08**
ω_2	0.0	0.0	-	0.0	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	**0.1**	0.0	0.0	-	0.0	0.0	0.0
ω_5	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0
ω_6	0.0	0.0	**0.06**	0.0	0.0	0.0	-	0.0
ω_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-

Multiple view motion-based recognition of gesture is commonly addressed by the complementary information processing paradigm relying on feature fusion and classification. For comparison purposes, we took the results provided in [4] where the alternative approach to multi-ocular recognition of gestures is analyzed. Similar error ratios are achieved but always relying on the assumption that only one individual is present in the scene and there are no occlusions.

V. 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 to perform classification.

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 [4], [6], [5].

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. Combination of motion detection together with a prior estimation of the body position of the person might allow a higher semantic analysis of the actions.

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

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

[4] A. F. Bobick and J. W. Davis, “The Recognition of Human Movement Using Temporal Templates,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 257–267, Mar. 2001.

[5] 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, July 2002.

[6] R. Rosales, “Recognition of Human Action Using Moment-Based Features,” Tech. Rep., Boston University, 1998.

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

[8] 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.

[9] 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.

[10] C.Canton-Ferrer, J.R.Casas, M.Tekalp, and M.Pardàs, “Projective Kalman Filter: Multiocular Tracking of 3D Locations Towards Scene Understanding,” *Lecture Notes on Computer Science*, vol. 3869, pp. 250–261, 2006.

[11] 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.

[12] J. L. Landabaso and M. Pardàs, “Foreground Regions Extraction and Characterization Towards Real-Time Object Tracking,” *Lecture Notes on Computer Science*, vol. 3869, pp. 241–249, 2006.

[13] G. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler, “A Real-Time System for Robust 3D Voxel Reconstruction of Human Motions,” in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2000.

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

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

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