2004 IEEE 6th Workshop on Multimedia Signal Processing

Invariant Action Classification with Volumetric Data

Fabio Cuzzolin, Augusto Sarti and Stefano Tubaro
Dipartimento di Elettronica e Informazione - Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano, Italy
Telephone: +39 02 2399 9640
Email: cuzzolin@elet.polimi.it, sarti@elet.polimi.it, tubaro@elet.polimi.it

Abstract— **We propose an action recognition algorithm in which the image sequences capturing a moving human body produced by a significant number of cameras are first used to generate a volumetric representation of the body by means of volumetric intersection. Classification is then performed directly on 3D data, making the system inherently insensitive to viewpoint dependence and motion trajectory variability. Suitable features are extracted from the voxset approximating the body, and fed to a hidden Markov model to produce a finite-state description of the motion. The Kullback-Leibler distance is finally used to classify new sequences.**

I. Introduction

Multi-camera systems have recently gained popularity in computer vision, thanks to a number of advantages that they exhibit over algorithms based on monocular views. Ambiguities in motion analysis due to perspective projection are resolved, and desirable properties like viewpoint invariance are inherently guaranteed. However, even if a few people have started to pose the problem in the volumetric context [1], [2], action recognition and activity detection algorithms are still largely based on 2D approaches, despite the fact that they can find more general and natural solutions in a multi-view setup. Recognition is in fact a complex task, as actions can be performed by different people in very different ways, with various speeds, and even the emotional state of the person can affect the evolution of the gesture.

In this paper we propose an action modeling and recognition approach in which images of the scene captured by a significant number of cameras are first used to generate a volumetric representation of a moving human body in terms of *voxsets*, by means of volumetric intersection. Recognition can then be performed directly on 3D data, allowing the system to avoid critical problems like viewpoint dependence and motion trajectory variability. We show how the use of appropriate local 3D features, inherently invariant with respect to trajectory variations, can significantly improve the performance of the classification (see also [3]).

One problem to consider is the so-called *time warping* issue: as actions may have different durations, a direct comparison between feature vectors at a given time is clearly impossible. *Hidden Markov models* ([4], [5]) have proven a quite successful method to cope with the matter. We adopt this formalism to model the action's dynamics from the collected 3D dataset, and the Kullback-Leibler distance between HMMs to classify new sequences.

Video surveillance problems [6] and activity detection for the implementation of "smart" environments are natural applications of the volumetric technique presented here.

II. 3D reconstruction and feature representation

A. Volumetric intersection

A simple but effective approach to volumetric reconstruction is the so-called *volumetric intersection* method, which exploits the silhouettes of an object extracted from all of its views.

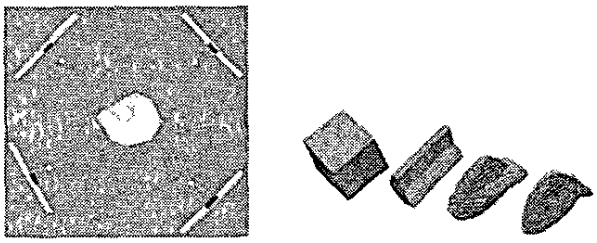

Fig. 1. Volumetric intersection. Left: the occlusion cones associated to the silhouettes of the body in each view are intersected, yielding a visual hull approximation of the actual object. Right: examples of reconstructions with respectively no views, a single view, several views.

The object is bound to be contained in the generalized cone generated by all the lines originating from the optical center of the camera and passing through the silhouette. It is then also contained in the intersection of all the corresponding "occlusion cones" (Figure 1-left). As Figure 1-right shows, the accuracy of the reconstruction critically depends on the number of viewpoints. The resulting *visual hull* will be the 3D reconstruction of the body.

A simple implementation of volumetric intersection starts from the discretization of the volume of interest into a *voxset* of reasonable size. Given a camera model and the associated calibration parameters, we then determine whether each voxel belongs to the object by checking whether it projects onto the inside of each of the available silhouettes.

B. Feature extraction

As voxsets are redundant descriptions of the body volume, we need to find a more concise representation (*feature*) of the

0-7803-8578-0/04/$20.00 ©2004 IEEE

moving body. We chose a rather simple description in terms of *bodypart positions*. We first estimate the motion direction of the person by interpolating the sequence of centers of mass $\bar{x}(t)$ along time by means of a *spline* (locally polynomial curve), and assuming as motion direction at time t the tangent to the interpolating curve in $\bar{x}(t)$. We then define as body reference frame at time t the triad $(\vec{d}(t), \vec{d}^{\perp}(t), \vec{z})$ where $\vec{z}$ is the vertical axes of the world reference frame, $\vec{d}(t)$ is the motion direction, and $\vec{d}^{\perp}(t)$ is its orthogonal complement in the xy plane of the world frame (Figure 2-left).

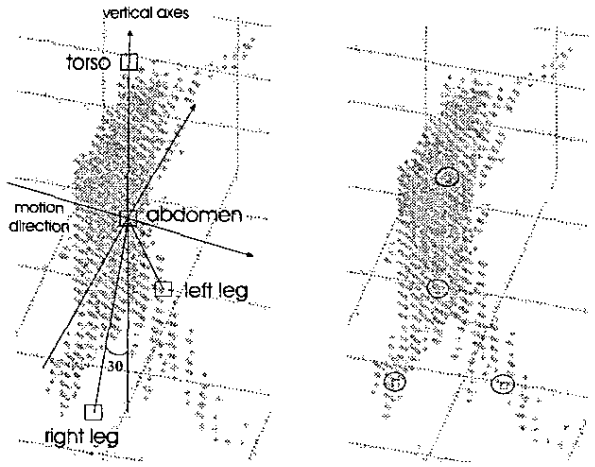

Fig. 2. Feature extraction. Body reference frame (left). Results of the 4-means clustering applied to the voxset (right).

Finally, to detect the bodyparts of the moving person we employ a *k-means clustering* algorithm with $n = 4$ clusters:

- in $t = 0$ the n cluster locations X_i, $i = 1, ..., n$ are assigned at random;
- given the cluster locations in $t = k$, a new set of centroids is achieved by
  - computing the distance $\|x - X_i\|$ between each point x of the voxset and each cluster location;
  - assigning each point x to the closer cluster;
  - computing a new cluster location as mean of the newly assigned points.

To guarantee the convergence of the four clusters to some desired positions (upper torso, abdomen, left and right leg) their initial positions in $t = 1$ are assigned to appropriate locations in the body reference frame (Figure 2-left). For $t > 1$ the old cluster positions in t are used as initial positions of the k-means algorithm in $t + 1$.

C. Linear discriminant analysis

Using the interpolated body trajectory to estimate the motion direction can be hazardous when the person halts or his/her motion is negligible. An alternative approach is estimating the frontal direction as the direction from which

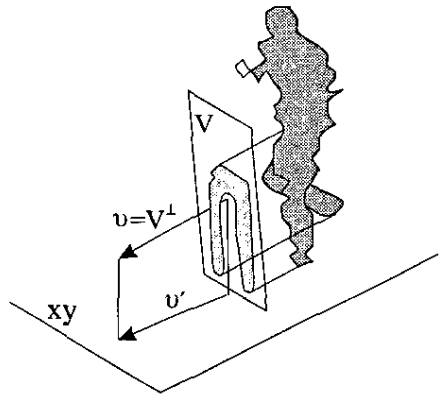

Fig. 3. LDA can be used to estimate the frontal view of the body by determining the maximal separation subspace of dimension 2.

we have the largest separation between the legs (exactly as humans do when asked to do the same). A mathematical tool called *linear discriminant analysis* provides us with a formal method to find the desired view.

Suppose that, given a dataset of N points x_j, $j = 1, ..., N$ in $\mathbb{R}^D$ with mean μ, partitioned in $\mathcal{K}$ classes $\mathcal{C}_k$ with mean μ_k, we wish to find the subspace in which the separation between the $\mathcal{K}$ clusters is maximum. This can be achieved through a linear transformation

$$y = W^T x \quad W = S_w^{-1} \cdot S_b$$

where the two matrices $S_b = \sum_{k=1}^{\mathcal{K}} N_k(\mu_k - \mu)(\mu_k - \mu)^T$ and $S_w = \sum_{k=1}^{\mathcal{K}} \sum_{x_j \in \mathcal{C}_k} (x_j - \mu_k)(x_j - \mu_k)^T$ are called *between-class* and *within-class* covariances respectively. W projects the dataset into a D-dimensional space. The target space's dimension can be reduced to d by selecting the first d eigenvalues. k-means can then be applied in this new space to detect the final clusters.

In our case, the legs' voxels can be processed by means of LDA, yielding the maximal separation subspace V with dimension $d = 2$. Quite obviously the frontal direction of the body will then be determined by the projection of the normal vector $v = V^{\perp}$ onto the floor plane xy in the world reference frame (Figure 3).

III. Action Modeling through HMMs

A *hidden Markov model* is a statistical model whose states $\{X_k\}$ form a *Markov chain*; the only observable quantity is a corrupted version y_k of the state called *observation process*. Using the notation in [7] we can associate the elements of the finite state space $\mathcal{X} = \{1, ..., n\}$ to coordinate versors $e_i = (0, .., 0, 1, 0, .., 0) \in \mathbb{R}^n$ and write the model as

$$\begin{cases} X_{k+1} = AX_k + V_{k+1} \\ y_{k+1} = CX_k + diag(W_{k+1})\Sigma X_k \end{cases}$$

where $\{V_{k+1}\}$ is a sequence of martingale increments and $\{W_{k+1}\}$ is a sequence of i.i.d. Gaussian noises $\mathcal{N}(0, 1)$. The

model parameters will then be the *transition matrix* $A = (a_{ij}) = P(X_{k+1} = e_i | X_k = e_j)$, the matrix C collecting the *means of the state-output distributions* (being the j-th column $C_j = E[p(y_{k+1}|X_k = e_j)]$) and the matrix Σ of the variances of the output distributions.

The set of parameters A, C and Σ of an HMM can be estimated, given a sequence of observations, through the *Expectation-Maximization* (EM) algorithm [7]

$$\{y_1, ..., y_T\} \mapsto A, C, \Sigma.$$

The likelihoods $\Gamma^i(y_{k+1})$ of the measurements y_{k+1} with respect to all the states $e_i, i = 1, ..., n$, are used to drive the recursive state estimation $\hat{X}_{k+1} = \sum_{i=1}^{n} A_i \langle \hat{X}_k, \Gamma^i(y_{k+1}) \rangle$, where n is the number of states, A_i is the i-th column of A and $\langle \cdot, \cdot \rangle$ is the usual scalar product.

Given a sequence of feature vectors extracted from the associated voxsets, EM yields as output a finite state representation of the motion, in which the transition matrix encodes the action's dynamics.

IV. Experimental results

For our tests we used a multi-camera TV studio at BBC R&D, located in Kingswood Warren, UK, equipped with a set of 12 calibrated, synchronized cameras placed in well separated positions around a studio of $4 \times 3.2 \times 2.5$ meters. As we were interested in action estimation in non-optimal conditions of acquisition, we selected $N = 5$ cameras covering the scene from a wide viewing angle.

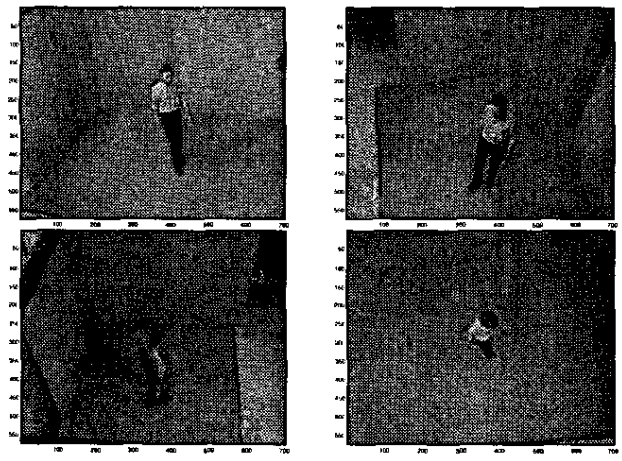

Fig. 4. An example of simultaneous views of a same studio scene. The person was asked to walk from one corner of the studio to the opposite one.

We then acquired 65 sequences, divided into three categories according the particular action performed: *"walk"*, *"walk and wave"*, and *"pick"* (an object from the ground). For each action category two different people were asked to perform several instances of these movements, following various trajectories and changing direction at will.

The BBC studio is equipped for a color segmentation of the acquired scene, yielding new frames in which only the object of interest is represented. The scene background was, in fact, covered by a special fabric that appears blue when illuminated by an appropriate light source. The desired segmentation is then done through multi-level thresholding of the chrominance channels, as these are much less sensitive to noise than the luminance channel. This chroma-keying process does not need to be too accurate, as the volumetric intersection takes care of removing most of the volumetric outliers. Once the sequence of silhouettes is produced, a sequence of volumetric reconstructions can be built through volumetric intersection (Figure 5). At each time step a feature vector is extracted as explained in Section II-B, so that a feature matrix is built for each sequence by collecting all the feature vectors along time, $y(t)$ for $t = 1, ..., T$. This feature matrix is then given as input to the EM algorithm, yielding the parameters of the HMM representing the action. We expected these feature vectors to be invariant with respect to nuisance parameters such as the trajectory followed, the size of the body, and the small "qualitative" differences between different people's movements. In fact, being the bodypart locations related to a reference frame associated to the motion direction, a person can walk along complex trajectories with no significant impact on the feature matrix. Figure 6 shows the large variability of the trajectories followed in the collected instances of the action "walk". Figure 7 instead compares two feature matrices associated with two of those walks performed by different people, showing a remarkable invariance.

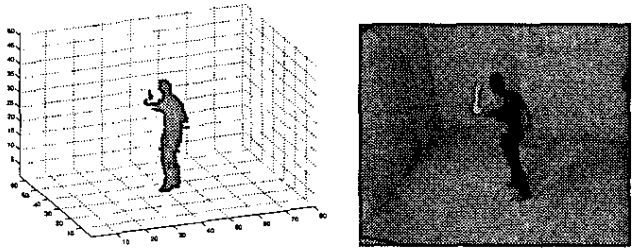

Fig. 5. Volumetric representation (left) associated with a real view (right).

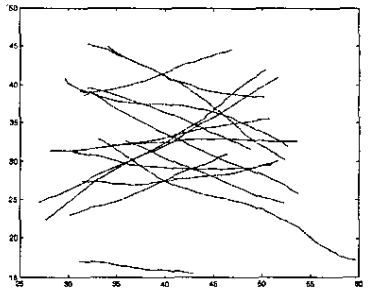

Fig. 6. Trajectories in the xy plane followed in the acquired instances of the "walk" action.

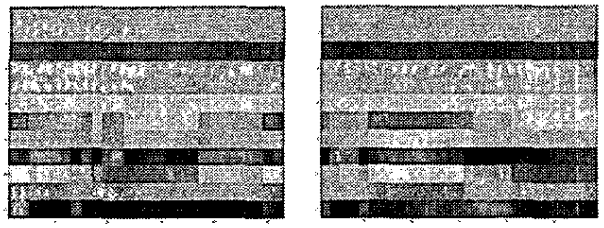

Fig. 7. Visual comparison between two feature matrices extracted from two distinct instances of the action "walk", performed by two different people in different directions. The matrices show the temporal evolution (horizontal axis) of the feature vectors extracted from the volumetric data.

Finally, Figure 8 shows the resulting hidden Markov model for the "walk" action. A model with 3 states proved to be adequate to represent this action, each state being associated with: the pose in which the left leg is extended; that in which both legs are aligned; and the one in which the right leg is thrust forward, respectively.

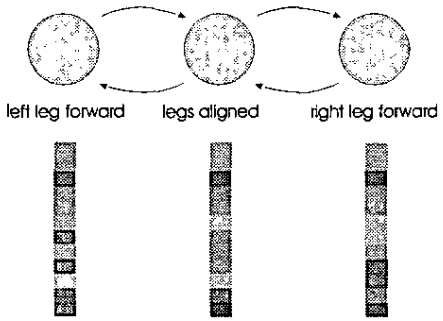

Fig. 8. Hidden Markov model associated with the "walk" action. The topology of the graph representing the action is given by the A matrix, while each state is represented by a feature vector c_j which is the j-th column of the matrix C.

Having built a model for each learned action category (walk, wave and pick), a new sequence can be classified by computing the associated model through the EM algorithm and directly comparing it to the learnt ones by means of the classical Kullback-Leibler distance [8]. The learnt models allowed to distinguish between instances of "walk" and "pick" (Figure 9), while the four-cluster representation of the body turned out to be inadequate to distinguish "wave" from "walk" when using a coarse volumetric representation. Nonetheless, even when using low-resolution voxsets, the system could still recognize "walk and wave" motions as instances of the "walk" action.

V. Perspectives

These first experiments prove how treating the action recognition task directly on 3D data is the most natural way of overcoming critical problems like viewpoint dependence, scale invariance, and other nuisance factors like trajectory variations. Problems like multi-body movements (for instance in automatic surveillance contexts) or occlusions are naturally

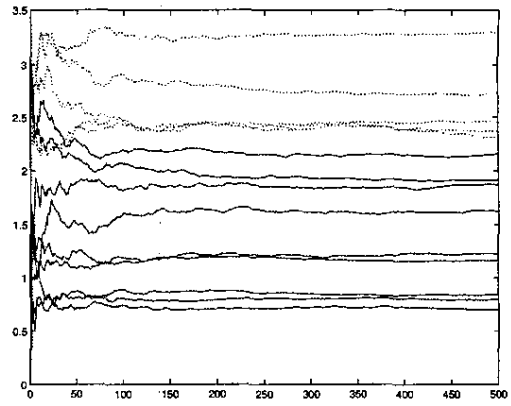

Fig. 9. Convergence of the Kullback-Leibler distance. The diagram plots against the number of steps of the K-L algorithm the distance between each model and a fixed HMM model chosen as basis. The K-L distances for instances of "walk" or "wave" are drawn as solid lines, while instances of the "pick" action are associated with dotted lines, showing a decent separation.

dealt with *before* the recognition stage. This motivates us to conduct more sophisticated analysis of the appropriate 3D feature representation, and study realistic situations in which the person performing the action is partially occluded from other objects, or shares the environment with other people. We are also currently conducting experiments with higher resolution voxsets, in order to detect arm motions through a six-cluster representation.

VI. Acknowledgements

The authors would like to thank the BBC R&D division for granting us permission to use of their studio, and Dr. Oliver Grau for his kind collaboration.

References

[1] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick, and A. Pentland, "Invariant features for 3-D gesture recognition," in *Proc. of FG'96*, 1996, pp. 157–162. [Online]. Available: citeseer.nj.nec.com/campbell96invariant.html

[2] S. Seitz and C. Dyer, "View-invariant analysis of cyclic motion," *International Journal of Computer Vision*, vol. 25, pp. 1–25, 1997.

[3] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Intl. J. of Computer Vision*, vol. 50(2), To appear, 2003. [Online]. Available: citeseer.nj.nec.com/rao02viewinvariant.html

[4] A. D. Wilson and A. F. Bobick, "Parametric Hidden Markov Models for gesture recognition," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21(9), Sept. 1999, pp. 884–900.

[5] M. Brand, N. Oliver, and A. Pentland, "Coupled HMM for complex action recognition," in *Proc. of Conference on Computer Vision and Pattern Recognition*, vol. 29, 1997, pp. 213–244.

[6] Y. Ivanov, C. Stauffer, A. Bobick, and E. Grimson, "Video surveillance of interactions," in *Proc. of the CVPR'99 Workshop on Visual Surveillance, Fort Collins, Colorado.*, November 1998. [Online]. Available: citeseer.nj.nec.com/ivanov99video.html

[7] R. Elliot, L. Aggoun, and J. Moore, *Hidden Markov models: estimation and control*, 1995.

[8] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *AT&T Technical Journal*, vol. Vol. 64(2), pp. 391–408, February 1985.