ACTION MODELING WITH VOLUMETRIC DATA

Fabio Cuzzolin, Augusto Sarti, Stefano Tubaro

Dipartimento di Elettronica e Informazione – Politecnico di Milano
Piazza Leonardo da Vinci 32, Milano, Italy
cuzzolin/sarti/tubaro@elet.polimi.it

ABSTRACT

In this paper we propose and test an action recognition algorithm in which the 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. The recognition stage is then performed directly on 3D data, allowing the system to avoid critical problems like viewpoint dependence and motion trajectory variability. Suitable features are extracted from the voxset representing the body, and fed to a classical hidden Markov model to produce a finite-state description of the motion.

1. 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], 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. The recognition stage 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 [2]).

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* [3, 4]) 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 propose the Kullback-Leibler distance between HMMs to classify new sequences.

Video surveillance problems [5] and activity detection for the implementation of "smart" environments are natural applications of the volumetric technique presented here.

2. 3D RECONSTRUCTION AND FEATURE REPRESENTATION

2.1. 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.

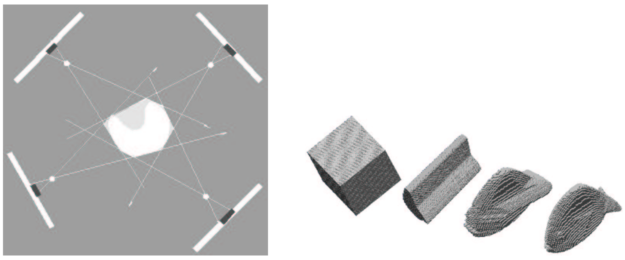


Fig. 1. Modeling through volumetric intersection. Left: the occlusion cones associated to the silhouettes of the body in each view are intersected, which results in a visual hull approximation of the actual object. Right: examples of reconstructions with respectively no views, one single view, several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

As the object is bound to be contained in the generalized cone generated by all lines that originate from the optical center of the camera and pass through the silhouette, it

This work has been partially funded by the FIRB "VICOM" project.

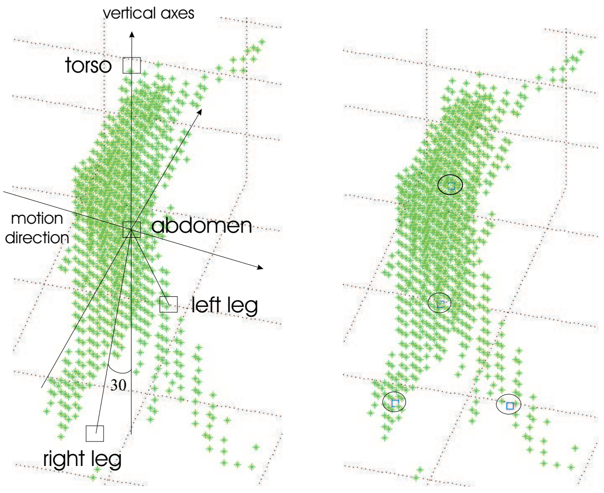
is also bound to be contained in the intersection of all the corresponding “occlusion cones” (Figure 1a). As Figure 1b shows, the accuracy of the reconstruction critically depends on the number of viewpoints. The resulting visual hull will be the 3D reconstruction of the body.

A simple and practical implementation of the volumetric intersection method starts from the discretization of the volume of interest into a *voxset* of reasonable size. We then determine whether each one of the voxels of the voxset belong to the object by simply checking whether that voxel projects onto the inside of each one of the available silhouettes, in accordance with the the adopted camera model and the available camera calibration parameters.

2.2. Feature extraction

As voxsets are redundant representations of the body volume, we need to adopt a more concise representation (*feature*) of the moving body. We considered a rather simple description in terms of *bodypart positions*.

First we estimate the motion direction of the person by interpolating the sequence of centers of mass $\bar{x}(t)$ along time using 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 $(\overrightarrow{d}(t), \overrightarrow{d}^{\perp}(t), \overrightarrow{z})$ where $\overrightarrow{d}(t)$ is the motion direction, $\overrightarrow{d}^{\perp}(t)$ its orthogonal direction in the xy plane, and $\overrightarrow{z}$ the vertical axes of the world reference frame (Figure 2a).



(a) (b)

Fig. 2. Feature extraction. a) Body reference frame. b) Results of the 4-means clustering applied to the voxset a).

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 are assigned at random;
- given the cluster locations in $t = k$, a new set of means 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) in $t = 1$ their initial positions are assigned to appropriate locations in the body reference frame (Figure 2a). For $t > 1$ the old cluster positions in t are used as initial positions of the k-means algorithm in $t + 1$.

Some problems could arise when the body speed is close to zero, since the motion direction estimation is not reliable. This can be fixed by adopting as x axes of the body frame the principal axes of the ellipsoid fitting the projection of the voxset onto the xy plane (Figure 3).

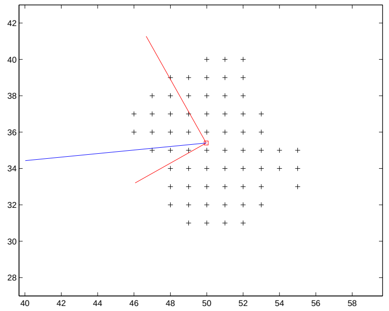


Fig. 3. View of the projection of a voxset onto the xy plane, together with the estimated motion direction and the axis of the ellipsoid which better fits the projected cloud.

3. 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 [6] 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 HMM 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 $C_j = E[p(Y_{k+1}|X_k = e_j)]$) and the matrix Σ of the variances of the output distributions.

A fundamental property of this class of models is its ability to self-learn the set of parameters A, C and Σ given a sequence of observations produced by the model through an application of the EM technique [6]:

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

In order to obtain the state estimate $\hat{X}_{k+1}$ associated to any new observation we compute the probabilistic distance $\Gamma^i(y_{k+1})$ between the measurement y_{k+1} and each state representative $C \cdot e_j$ in the observation space

$$\hat{X}_{k+1} = \sum_{i=1}^{n} A_i \langle \hat{X}_k, \Gamma^i(y_{k+1}) \rangle$$

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.

4. 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 are interested in action estimation in non-optimal conditions of acquisition, we selected $N = 5$ cameras, covering the scene from a wide viewing angle.

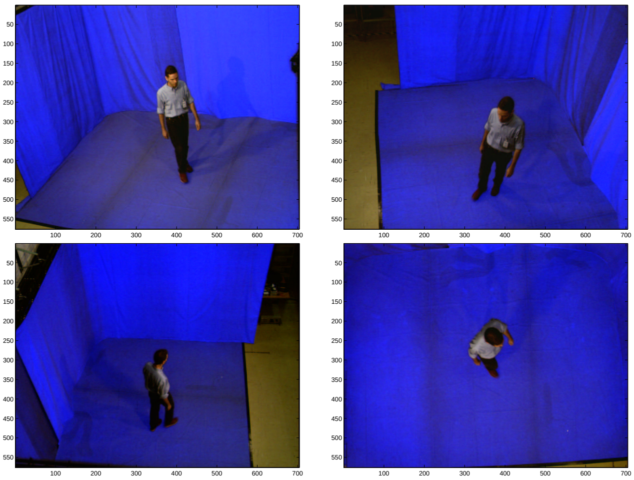


Fig. 4. An example of simultaneous views of a same studio scene. The person was asked to walk from one corner of the studio to the opposite one.

For these experiments we acquired 65 sequences, divided into three categories according the particular action performed: "*walk*", "*walk and wave*", and "*pick*" and object from the ground. For each category of actions two different people performed several instances of the motion, following various trajectories and even changing direction in the middle of the action.

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. Each one of the 12 cameras is equipped with such a projector. 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 accuare as the volumetric intersection will take care of removing most of the volumetric outliers (see Figure 5).

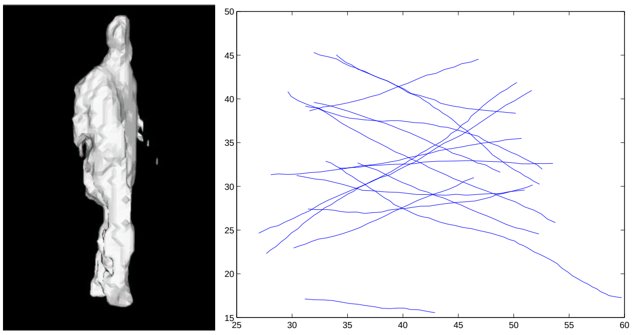


Fig. 5. Volumetric representation of the person represented in Figure 4 (left). Trajectories in the xy plane followed in the instances of the "walk" action (right).

Once the sequence of silhouettes is produced, a sequence of volumetric reconstructions can be built through volumetric intersection. At each time step a feature vector is extracted as explained in Section 2.2, 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 an EM algorithm, which computes 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 chosen by the person, the size of the body, and the small "qualitative" differences between walking or waving performed by different people. In fact, being the bodypart locations related to a reference frame associated to the motion direction, a person can walk in complex curved trajectories with no significant impact on the feature matrix. Figure 5b shows the large variability of the trajectories followed in instances of the action "walk". Figure 6 instead compares two feature matrices associated to two of those walks performed by different people, showing a remarkable invariance.

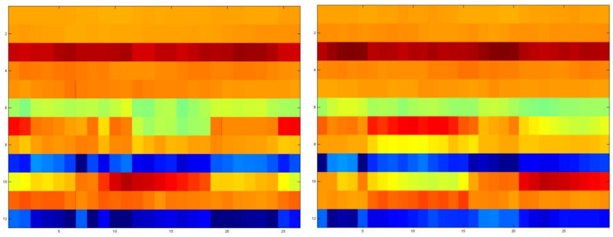


Fig. 6. Visual comparison between two feature matrices extracted from two distinct instances of the action "walk", performed by two different people in different directions. The matrices show the temporal evolution (horizontal axis) of the feature vectors extracted from the volumetric data.

Finally, Figure 7 shows the hidden Markov model of the "walk" action. Given the feature matrices of Figure 6, a model with 3 states proved to be suitable to represent this action, each state being associated to: 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.

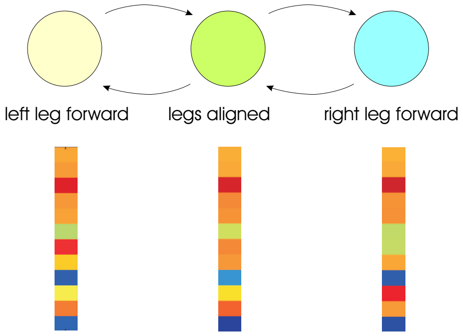


Fig. 7. Hidden Markov model associated to the "walk" action. The topology of the graph representing the action is given by the A matrix, while each state is associated to a feature vector c_j which is the j-th column of the matrix C.

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 by directly comparing it to the learnt ones by means of the classical Kullback-Leibler distance [7]. The models we obtained proved good enough to distinguish between instances of "walk" and "pick". The four-cluster representation of a body, on the other hand, does not allow us to distinguish between "wave" and "walk" when using a coarse volumetric representation (the arms of the person require high spatial resolution). Nonetheless even when using low-resolution voxsets, the systems can still recognize the "walk and wave" action as an instance of the "walk" action.

5. 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 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 currently conducting experiments with higher resolution voxsets in order to detect arm motion through a six-cluster representation.

6. ACKNOWLEDGEMENTS

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

7. 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.

[2] 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.

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

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

[5] 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.

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

[7] 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.