Automatic Discovery of Action Taxonomies from Multiple Views

Daniel Weinland * Remi Ronfard Edmond Boyer
Project PERCEPTION, INRIA Rhone-Alpes,
38334 Montbonnot Saint Martin, France
{weinland, ronfard, eboyer}@inrialpes.fr

Abstract

We present a new method for segmenting actions into primitives and classifying them into a hierarchy of action classes. Our scheme learns action classes in an unsupervised manner using examples recorded by multiple cameras. Segmentation and clustering of action classes is based on a recently proposed motion descriptor which can be extracted efficiently from reconstructed volume sequences. Because our representation is independent of viewpoint, it results in segmentation and classification methods which are surprisingly efficient and robust. Our new method can be used as the first step in a semi-supervised action recognition system that will automatically break down training examples of people performing sequences of actions into primitive actions that can be discriminatingly classified and assembled into high-level recognizers.

1 Introduction

Recognizing actions of human actors from video is an important topic in computer vision with many fundamental applications in video surveillance, video indexing and social sciences. From a computational perspective, actions are best defined as four-dimensional patterns in space and in time [10]. Yet, much current research in computer vision ignores this fact and attempts to learn action models directly from monocular video [3, 6, 1]. In our work, we use multiple video cameras and shape-from-silhouette techniques to obtain four-dimensional recordings of action sequences. We compute new motion descriptors based on - *motion history volumes* - which fuse action cues, as seen from different viewpoints and over short time periods, into a single three dimensional representation. From that representation, we are able to segment the action streams into primitives and to cluster those primitives into a hierarchy of primitive action classes.

Our long-term goal is to automatically generate high-level descriptions of video sequences in terms of the actions that can be recognized or inferred from the given visual input. Actions generally fall under two distinct categories - composite actions which can be broken down into distinct temporal parts or segments, and primitive actions, which cannot be broken down further. In order to build a general action recognizer, we need the ability to break down a given sequence into primitive action segments, to label those segments into primitive actions using a vocabulary of learned action models, and to assemble the labeled segments into composite actions using concept hierarchies [8] or grammars [11].

In this work, we use a novel motion descriptor based on the *motion history volume* (MHV) which summarizes the action content of a short multi-view sequence without knowledge of body parts [15]. We automatically segment action sequences into primitive actions which can be represented by a single MHV and we cluster the resulting MHVs into a hierarchy of action classes, which allow us to recognize multiple occurrences of repeating actions. We are able to perform those two steps automatically, mainly because MHVs work in a volume space which considerably reduces the ambiguities traditionally associated with changes in viewpoints and occlusions even in multiple views.

As a concrete example, we asked two members of our lab to perform a sequence of simple actions, each repeated several times with different poses and styles, in front of 6 calibrated cameras. The resulting data set consists of unsegmented and unlabeled synchronized video sequences such as the one depicted in Figure 1. Using the new motion descriptor, we were able to segment (Section 5) and cluster (Section 6) such sequences into primitive actions, which we used as training examples for learning statistical classifiers. Such a semi-supervised scheme is important in practical terms because it facilitates the creation of large training sets for action recognition in the large.

Our method generates action taxonomies based on purely visual cues since we create higher-level action classes by abstracting two or more recorded actions which
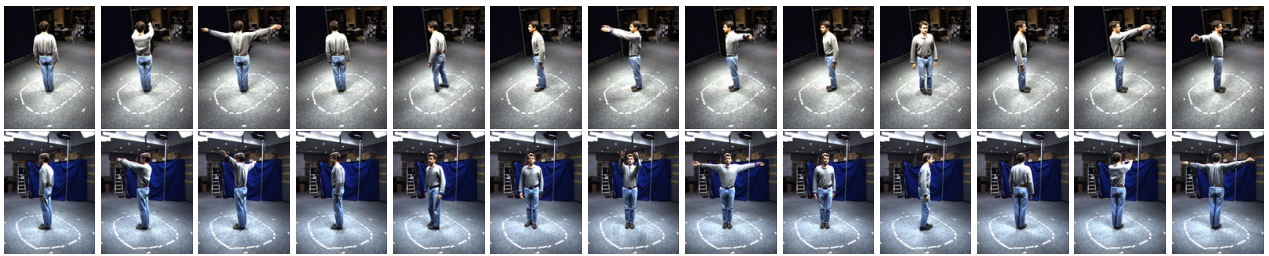
Figure 1. Example action sequence: Raise arms - rotate arms - turn left - raise arms - rotate arms - turn left - raise arms - rotate arms, seen from two different viewpoints. Such sequences are difficult to segment and label consistently from monocular cues, but are easily segmented and labeled using our view-independent motion descriptors.

look the same from all viewpoints (as measured by the differences in a metric space of motion descriptors extracted from their MHVs). We believe this is an important step towards building complete, semantic taxonomies of actions and plans.

The paper is organized as follows. We review related work in Section 2. We briefly review motion history volumes and associated view-independent motion descriptors in Sections 3 and 4. We describe our segmentation algorithm in Section 5 and our clustering algorithm in Section 6. Both algorithms are based on the motion descriptors introduced in Section 4. Finally in Section 7 we describe a semi-supervised action classification system which uses the proposed algorithms to automatically segment and label the training and test sequences, and report initial results obtained on a limited but realistic data set.

2   Related work

Segmentation and labeling of action sequences from *monocular video* is a difficult problem that has received considerable attention in recent years. Rittscher et al. learn dynamical models of actions from tracked contours and use them to segment new sequences [12]. Rui and Anandan perform an SVD decomposition of a long sequence of optical flow images and detect discontinuities in the trajectories of selected SVD components to segment video into *motion patterns* [13]. Zelnik-Manor and Irani cluster video sequences into *events* using normalized cuts on multiresolution sequences of spatio-temporal gradient magnitudes [16]. Brand and Kettnaker use unsupervised HMMs to perform simultaneous segmentation and clustering of actions from sequences of human silhouettes [3]. Wang et al. also use unsupervised HMMs to segment 2D hand motions and extract a vocabulary of musical conducting gestures, which allows them to describe video sequences optimally in the sense of minimum description length [14]. Feng and Cham compare methods for segmenting action sequences with or

without body part correspondences and propose a hybrid scheme that can handle *ambiguous correspondences* [5]. All such methods work only with restricted variations in viewpoint, which make them ill-suited to cases such as of Figure 1 where each action is performed multiple times with vastly different poses.

Segmentation and labeling of action sequences from *multiple views* is a relatively little-studied area. Previous work assumes either that the cameras are uncalibrated (so that reconstruction is not possible) or that a full human body model can be recovered (so that reconstruction includes body part recognition and tracking). Thus, Marr and Vaina discuss the problem of segmenting the 3D movement of shapes and suggest the use of local minima of the 3D motion of human limbs as natural transitions between primitive movements [9]. Campbell et al. investigate several view-invariant features for action classification from face and hand tracking based on using multi-view stereo [4]. Davis and Bobick use *motion templates* in multiple views, but they assume uncalibrated cameras and are therefore unable to perform 3D reconstruction [2]. Similarly, Ogale et al. cluster action sequences in multiple views separately by detecting minima and maxima of optical flow inside silhouettes and matching the selected silhouettes using phase correlation [11]. This allows them to learn action grammars from examples recorded by multiple cameras, but their grammars remain viewpoint-dependent.

To the best of our knowledge, no previous work has attempted to perform segmentation and clustering from *volumetric* reconstructions. In this paper, we propose such a method, which extends monocular methods most naturally by introducing view-invariant motion descriptors built from silhouettes in multiple calibrated views. Compared with previous work, our method has the advantage that we perform all three steps of segmenting, clustering and classifying action sequences in 3D with a representation which is fully view-invariant, and is much simpler to recover than a full human body model.
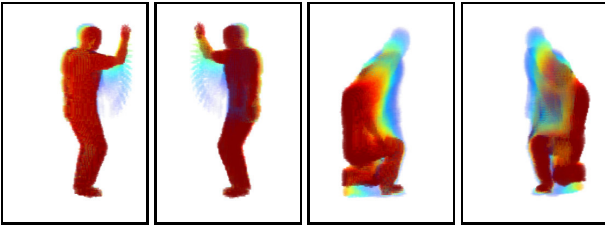
Figure 2. Example motion history volumes: "Lift arm" and "knee", rendered from different viewpoints. Colors: red = current, ..., blue = maximum duration, encode time of last occupancy.

3 Motion History Volumes

In this section, we present the 3D motion templates on which we ground our approach. These templates are a 3D generalization of the 2D motion templates introduced by Bobick and Davis in [2]. Both 2D and 3D templates are based on image silhouettes, which are binary valued functions indicating object occupancies in image projections.

Motion templates encode the history of motion occurrences. In 2D images, pixel values are therefore multiple-values recording how recently motion occurred at a pixel. The extension to 3D is straightforward by considering voxels instead of pixels, and the space occupancy function $D(x, y, z, t)$ over time steps t. Voxel values in the MHV at time t are then defined by:

$$v_\tau(x, y, z, t) = \begin{cases} \tau & \text{if } D(x, y, z, t) \\ \max(0, v_\tau(x, y, z, t-1) - 1) & \text{oth.} \end{cases} \quad (1)$$

where τ is the maximum duration a motion is stored.

The input occupancy function $D(x, y, z, t)$ is estimated using silhouettes and is defined by the visual hull at time t. Voxelic visual hulls are easy to compute and yield robust 3D representations. Note however that, as for 2D motion templates, different body proportions may still result in different templates. Figure 2 shows examples for motion history volumes.

4 Motion Descriptors

To compare or discriminate motions, we need to find a representation which is invariant to transformations in locations, orientations or sizes. To this purpose, we use both alignment and invariant descriptors based on motion templates and Fourier transform. The idea is first to center scale-normalized motion history volumes into a cylindrical coordinate system where the z-axis is aligned with the vertical direction. Hence, dependencies on scale and horizontal translations are removed. For rotations around the vertical

axis, we use the fact that they correspond to translations in the cylindrical coordinate systems, and that a function $f_0(x)$ and its translated counterpart $f_t(x) = f_0(x - x_0)$ only differ by a phase modulation after Fourier transform:

$$F_t(k) = F_0(k)e^{-j2\pi kx_0}. \quad (2)$$

Thus absolute values of the Fourier transform are rotation invariant descriptors.

The choice made here is motivated by the assumption that similar actions only differ by rigid transformations composed of scale, translation, and rotation around the z-axis. Of course, this does not account for all similar actions of any body shape, but it appears to be reasonable in most situations. In addition, restricting the Fourier-space representation to the lower frequencies also implicitly allows for additional degrees of freedom in object appearances and action executions. Our experiments also show that Fourier magnitudes provide more discriminative information than correlation features when comparing actions. The following section details our exact implementation.

Alignment   We express the motion templates in a cylindrical coordinate-system:

$$v(\sqrt{x^2 + y^2}, \tan^{-1}\left(\frac{y}{x}\right), z) \to v(r, \theta, z).$$

Thus rotations around the z-axis results in cyclical translation shifts:

$$v(x\cos\theta_0 + y\sin\theta_0, -x\sin\theta_0 + y\cos\theta_0, z) \to v(r, \theta + \theta_0, z).$$

We center and scale-normalize the templates. In detail, if v is the volumetric cylindrical representation of a motion template, we assume all voxels that represent a time step, i.e. for which $v_0(r, \theta, z) > 0$, to be part of a point cloud. We compute the mean μ and variances σ_r and σ_z in z- and r-direction. The template is then shifted, so that $\mu = 0$, and scale normalized so that $\sigma_z = \sigma_r = 1$. We choose to normalize in z and r direction, instead of a PCA based normalization, focusing on the main directions human differ on, and assuming scale effects dependent on positions to be rather small. This method may fail aligning e.g. a person spreading its hand with a person dropping its hand, but gives good results for people performing similar actions, which is more important.

Invariant descriptors   In the new coordinate system we apply a 1D Fourier-transform over θ for each value r and z:

$$V(r, k_\theta, z) = \int_{-\pi}^{\pi} v(r, \theta, z)e^{-j2\pi k_\theta \theta} d\theta, \quad (3)$$

and take as invariant features the magnitudes:

$$f(r, k_\theta, z) = |V(r, k_\theta, z)|. \quad (4)$$

Note that various combination of the Fourier transform could be used here, for example magnitudes of the 3D Fourier-transform over all dimensions r, θ, z as we did in [15]. We use the Fourier transform over the single dimension θ to preserve exact spatial information in the remaining directions. Such spatial information appears to be important when segmenting motions into elementary actions. The counterpart is that the above descriptor (4) is ambiguous with axial symmetries along the z-axis, hence similar actions performed by the left or right body parts can be difficult to discriminate.

It should also be mentioned here that to preserve the properties of the Fourier transform (e.g. robustness to noise, separation in fine and coarse features) for all dimensions, an additional 2D Fourier-transform can be applied to $f(r, k_\theta, z)$ for r and z:

$$\hat{V}(\omega_r, k_\theta, \omega_z) = \iint_{-\infty}^{\infty} |V(r, k_\theta, z)| e^{-j2\pi(\omega_r r + \omega_z z)} dr dz. \quad (5)$$

5 Temporal Segmentation

Temporal segmentation consists in splitting a sequence of motions into elementary segments. It is a necessary preliminary step to higher level processing of motion sequences including classification and clustering. In supervised approaches, segments are usually manually labeled in an initial set of motion sequences, and further operations are achieved by correlating unknown motion sequences with these learned segments on a frame by frame basis, using possibly various temporal scales [2, 7]. In this paper, we do not assume such *a priori* knowledge and propose instead a simple but efficient approach to automatically segment 3D motion sequences.

Any temporal segmentation relies on the definition of elementary motion segments. There are two main approaches to segmentation: Energy minima can be used to detect reversal of motion direction, following an early proposal by Marr and Vaina [9]. Or discontinuities can be used to detect changes in the temporal pattern of motion [13]. From experiments we found energy minima more stable, i.e. similar action sequences are segmented more consistently. The function over time that we segment is then a global motion energy function. This function is an approximation of the global body velocity estimated using the motion history volumes. It is based on the observation that rest states correspond to instants where few motions only occur, and thus result in few voxels encoding motion in the MHV, when small temporal windows are considered. Therefore, segment detection simply consists in finding minima of the sum of voxel values in the MHV, assuming a small value for τ in 1. Figure 3 shows several examples of sequences segmented this way. As can be seen in the figure, detection of energy minima is fairly unambiguous in this examples.
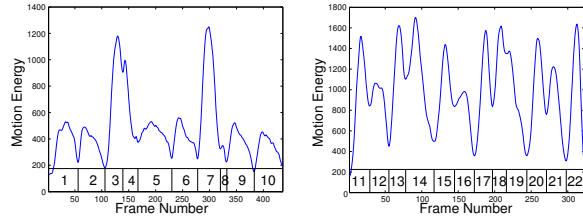
Figure 3. Motion Energy for action: Lift arms - rotate arms - lower arms and turn in new position. Executed three times by (left) female actor, (right) male actor. Local energy minima serve as segmentation criteria of sequences. Motion volumes for each segment are shown in Figure 4.

In the implementation we use a derivative of Gaussian filter and zero crossing to detect the minima. Parameter τ in equation (1) was set to constant 10 frames during all experiments. Temporal scale was not important for detection of all relevant segments. In practice, the minima detection appears to be very successful in segmenting motions, even for coupled motions, like moving torso and arms in parallel, local minima occur. Of course, this measure is still sensitive to small variations of velocity that can result in local minima. However, by allowing a possible over-segmentation the method will detect most of the motion segment boundaries.

6 Action taxonomies

Given a segmented action sequence, we would like to recognize multiple occurrences of the same primitive actions and to label the sequence accordingly. This capability will be important in the next section when we attempt to train classifiers for all primitive actions in a semi-supervised fashion.

We build an action taxonomy from a segmented sequence by hierarchically clustering the segments into classes. Initially, each segment is a single occurrence of its own action class, and is represented as a single point in the space of view-invariant motion descriptors of Section 4, which is a high-dimensional Euclidean space. We then apply a standard hierarchical clustering method to the segments. This creates a binary tree of action classes, where each class is now represented by a point cloud in the space of motion descriptors (see Figure 6).

We report experiments on two different datasets of increasing complexity. In each we segment the sequences as explained in section 5 and compute a single MHV per segment. This is illustrated in Figures 4 and 8. The experiments
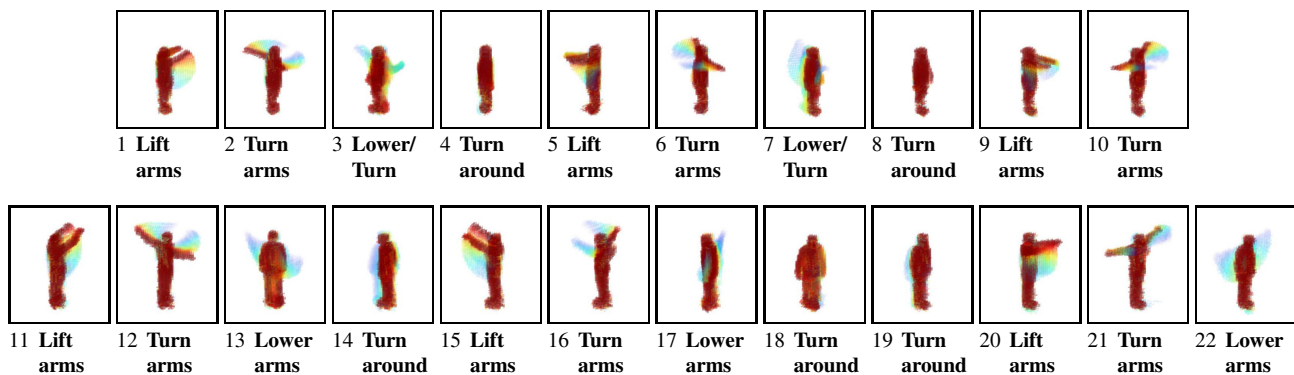
Figure 4. History volumes computed at segments of varying duration, and their clusters, using segmentation from Figure 3. (Top) female actor repeating three times: Lift arms ahead - rotate arms - lower arms and turn in new position. (Bottom) the same done by a male actor, from original sequence shown in Figure 1. The clusters are labeled manually for presentation purposes.

were conducted on MHVs obtained from 6 silhouettes extracted using a standard background subtraction method. The resulting motion templates were mapped into a discrete cylindrical coordinate representation of size $64 \times 64 \times 64$. Clustering was achieved using an agglomerative scheme, where the distance between objects is the Euclidean distance, and clusters were linked according to their furthest neighbor. The first dataset shows how actions performed by different persons, with different bodies, are handled by our system. The second dataset is a more realistic set of natural actions in arbitrary orders. Its interpretation is less straightforward, but it gives strong insights on the potential of our motion descriptors to yield consistent high-level interpretations.

6.1 Clustering on Primitive Actions

Here a dataset of 22 motion sequences performed by both a male and a female actor were considered. Segmented key actions are shown in Figure 5. The actors perform successively each action three times while changing their orientations in between. The automatic motion segmentation returns 203 motion volumes (100 for the woman, 103 for the man). We start by computing a dendrogram of all male segments, using Euclidean distances and furthest neighbor assignments. A good trade-off between motion variation within single clusters and multiple clusters having same labels is then to cut the hierarchy into 21 clusters. All segments inside these clusters are labeled according to the most obvious interpretation. From these labels, the 21 clusters are then labeled with respect to the most current actions which occurs in each cluster. Figure 6 shows the labeled dendrogram. Within these clusters, 7 (6.8%) actions were obviously assigned a wrong cluster, 4 actions give birth to
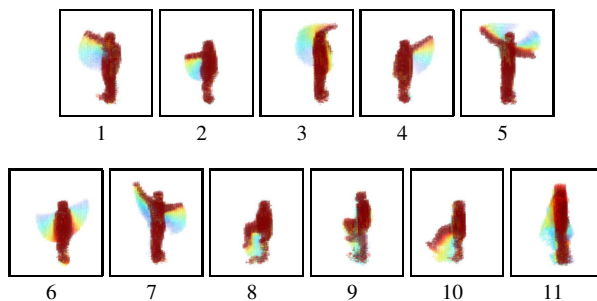

Figure 5. Perspective views of the motion history volumes computed for each action category. (1) lift right arm ahead. (2) lift right arm sideways. (3) lift left arm sideways ahead. (4) lift left arm sideways. (5) rotate both arms lifted. (6) lower both arms sideways. (7) lift both arms sideways. (8) lift right leg bend knee. (9) lift left leg bend knee. (10) lift right leg firm. (11) jump.

single clusters, and one cluster is ambiguous (lower or lift arm sideways).

We next compute a hierarchy from the male and female data. The procedure is the same as in the previous experiment. Du to higher variations in the dataset the clusters result in a coarser action grouping. A good trade-off between motion variation within single clusters and multiple clusters having same labels is this time to cut the hierarchy into 9 clusters, as shown in Figure 7. With respect to this labeling only two actions are wrongly assigned.
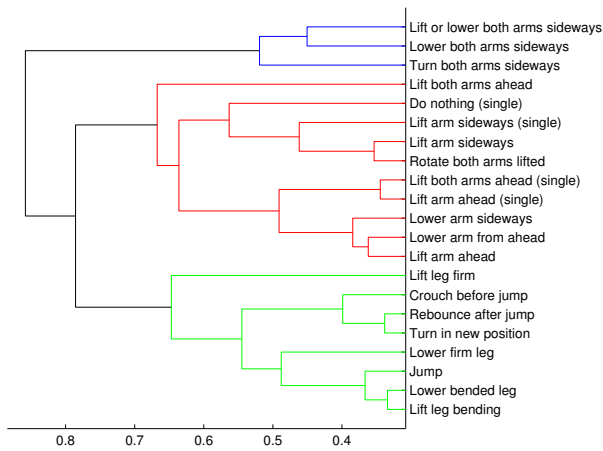
Figure 6. Hierarchical clustering of 103 male actions. 21 top nodes labeled with respect to the most occurring action.
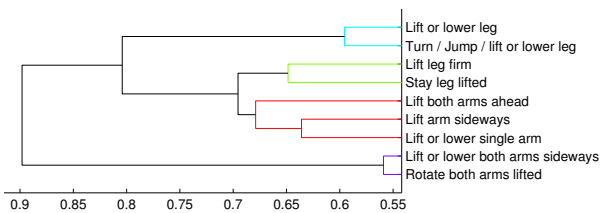

Figure 7. Hierarchical clustering of 203 male and female actions. 9 top nodes labeled with respect to the most occurring action.

6.2 Clustering on Composite Actions

In another clustering experiment we used a different dataset of actions with a much more complex semantics. Those sequences are pantomimes of various daily life actions such as catching a ball, picking up, stretching, laughing, etc. The segmentation and clustering methods were applied to each of these sequences. Figures 8 and 9 show the segmented motion templates and the hierarchy obtained for one such sequence. Again groups of higher level actions in Figure 9 have a simple interpretation such as lift or lower arms. Note also the group *rest in position* where segments without motion, typically between actions, have been consistently clustered.

7 Semi-supervised classification

In this section we use the MHV clusters as training data to learn discriminant classifiers for each of the discovered action classes. We use the motion templates that have been
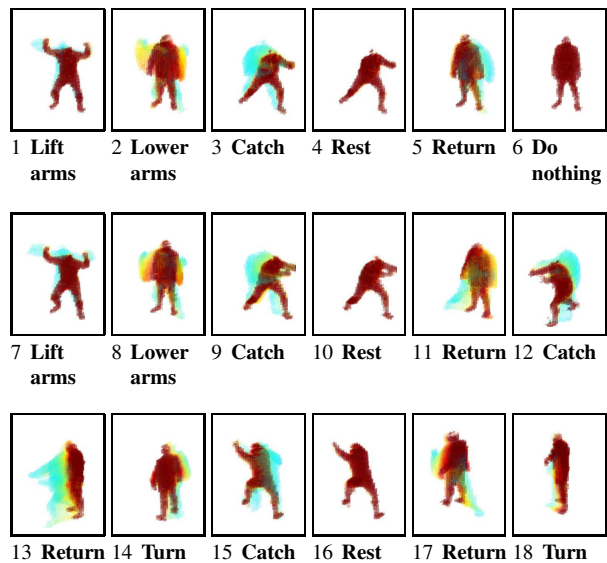

Figure 8. History Volumes for pantomime sequence "catching ball".

automatically extracted in the previous section, i.e. we use the 11 actions corresponding to the key labels in Figure 5. Each action is represented each by 3 samples per actor. One example per class is shown in Figure 5.

We split the set into two configurations: woman/man and man/woman. While simple, this test shows how the proposed descriptors discriminate actions with different bodies. Every action class in the data-set is represented by the mean value of the descriptors over the available population in the action training set. Any new action is then classified
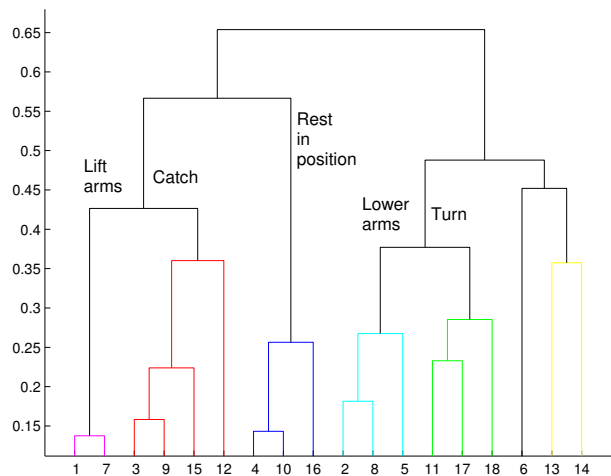
Figure 9. Hierarchical clustering of "catching ball" sequence.
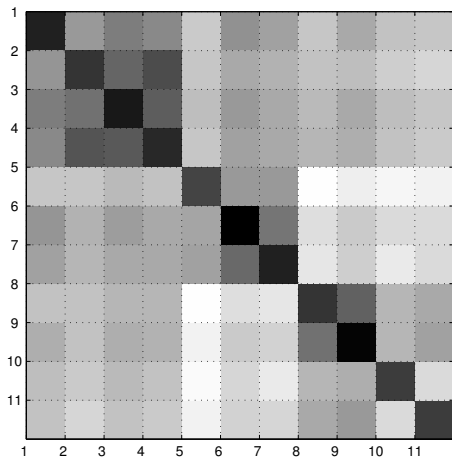
Figure 10. Average distances in feature space between male action classes and female samples. Actions see Figure 5.

according to a Mahalanobis distance associated to a PCA (Principal Component Analysis) based dimensional reduction of the data vectors.

One pooled covariance matrix Σ based on all training samples $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ was computed:

$$\Sigma = \frac{1}{n} \sum_{i}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top, \qquad (6)$$

where $\mathbf{m}$ represents the mean value over all training samples.

The Mahalanobis distance between feature vector $\mathbf{x}$ and a class mean $\mathbf{m}_i$ representing one action is:

$$d(\mathbf{m}_i, \mathbf{x}) = (\mathbf{x} - \mathbf{m}_i)^\top V \Lambda^{-1} V^\top (\mathbf{x} - \mathbf{m}_i),$$

with Λ containing the k largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, and V the corresponding eigenvectors of Σ. Thus feature vectors are reduced to k principal components.

In all tests x are the vectorized $6 \times 6 \times 6$ lowest complex valued frequencies of equation (5), that are further reduced to the 32 largest principal components. Independent whether the classes are learned from the male/female data, we achieve in both cases a classification rate of 100%. A confusion matrix of average distances is shown in Figure 10, with surprisingly good results even with respect to axial symmetry.

8 Conclusion

In this paper, we have introduced new methods for segmenting and clustering sequences of volumetric reconstructions of a human actor performing actions, without recognition or tracking of body parts. This has allowed us to learn classifiers for a small vocabulary of primitive actions, independently of style, gender and viewpoint. We have also applied our algorithms to discover meaningful hierarchies of action concepts in more complex composite sequences. We are currently using our new semi-supervised method to build training sets with more actions, actors and styles. In future work, we plan to use those techniques to learn statistical models of composite actions by simultaneously learning the component actions and their grammars.

References

[1] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *ICML*, 2004.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001.

[3] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *PAMI*, 22(8):844–851, August 2000.

[4] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, and A. Pentland. Invariant features for 3-d gesture recognition. In *FG*, pages 157–163, October 1996.

[5] Z. Feng and T. Cham. Video-based human action classification with ambiguous correspondences. In *V4HCI*, 2005.

[6] R. Green and L. Guan. Quantifying and recognizing human movement patterns from monocular video images. *TCSV*, 14, February 2004.

[7] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, October 2005.

[8] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 50(2):171–184, 2002.

[9] D. Marr and L. Vaina. Representation and recognition of the movements of shapes. *Proc. R. Soc. B*, 214:501–524, 1982.

[10] J. Neumann, C. Fermüller, and Y. Aloimonos. Animated heads: From 3d motion fields to action descriptions. In *DE-FORM/AVATARS*, 2000.

[11] A. Ogale, A. Karapurkar, and Y. Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *WDV*, 2005.

[12] J. Rittscher, A. Blake, A. Hoogs, and G. Stein. Mathematical modelling of animate and intentional motion. *Phil. Trans. R. Soc. B*, pages 475–490, 2003.

[13] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *CVPR*, pages 1111–1118, 2000.

[14] T. Wang, H. Shum, Y. Xu, and N. Zheng. Unsupervised analysis of human gestures. In *PCM*, pages 174–181, 2001.

[15] D. Weinland, R. Ronfard, and E. Boyer. Motion history volumes for free viewpoint action recognition. In *PHI*, 2005.

[16] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *CVPR*, December 2001.