Motion History Volumes for Free Viewpoint Action Recognition

Daniel Weinland, Rémi Ronfard, Edmond Boyer

Motion History Volumes for Free Viewpoint Action Recognition

Daniel Weinland* Remi Ronfard Edmond Boyer

Project MOVI, INRIA Rhone Alpes,
Montbonnot Saint Martin, France-38334.

Abstract

Action recognition is an important and challenging topic in computer vision, with many important applications including video surveillance, automated cinematography and understanding of social interaction. Yet, most current work in gesture or action interpretation remains rooted in view-dependent representations. This paper introduces Motion History Volumes *(MHV) as a free-viewpoint representation for human actions in the case of multiple calibrated, and background-subtracted, video cameras. We present algorithms for computing, aligning and comparing MHVs of different actions performed by different people in a variety of viewpoints. Alignment and comparisons are performed efficiently using Fourier transforms in cylindrical coordinates around the vertical axis. Preliminary results indicate that this representation can be used to learn and recognize basic human action classes, independently of gender, body size and viewpoint.*

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 [15]. Video recordings of actions can similarly be defined as three-dimensional patterns in image-space, and in time, resulting from the perspective projection of the world action onto the image plane at each time instant. Recognizing actions from a single video is plagued with the unavoidable fact that parts of the action are hidden from the camera because of self-occlusions. That the human brain is able to recognize actions from a single viewpoint should not hide the fact that actions are firmly four-dimensional, and, furthermore, that the mental models of actions supporting recognition may also be four-dimensional.

*D. Weinland is supported by a grant from the EC under the EST Marie-Curie Project Visitor.

In this paper, we investigate how to build spatio-temporal models of human actions that can support categorization and recognition of simple action classes, independently of viewpoint, actor gender and body sizes. Action recognition can be separated in two separate tasks. The first task is the extraction of motion descriptors from visual input, and the second task is the classification of the descriptors into various levels of action classes, from simple gestures and postures to primitive actions to higher levels of human activities [12]. That second task can be performed by learning statistical models of the temporal sequencing of motion descriptors. Popular methods for doing this are hidden markov models and other stochastic grammars [9]. In our work, we focus on the extraction of motion descriptors from multiple cameras, and their classification into *primitive actions* such as raising and dropping hands and feet, sitting up and down, jumping, etc. To this aim, we introduce 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.

Previous work on motion descriptors uses positions and velocities of human body parts [6], but such information is difficult to extract automatically during unrestricted human activities. Motion descriptors which can be extracted automatically, and have been used for action recognition, are optical flows [4], motion templates [2], and space-time volumes [21, 22]. Such descriptors are not invariant to viewpoint, which can be partially resolved by multiplying the number of action classes by the number of possible viewpoints [2], relative motion directions [4], and point correspondences [21, 22]. This results in a poorer categorization and an increased complexity.

In this research, we investigate the alternative possibility of building free-viewpoint class models from view-invariant motion descriptors. The key to our approach is the assumption that we need only consider variations in viewpoints around the central vertical axis of the human body. Within this assumption, we propose a representation based on Fourier analysis of motion history volumes in cylindrical coordinates (see figure 1). Such a representation fits nicely within the framework of Marr's 3D model [14] which has been advocated as a useful tool for representing action cat-

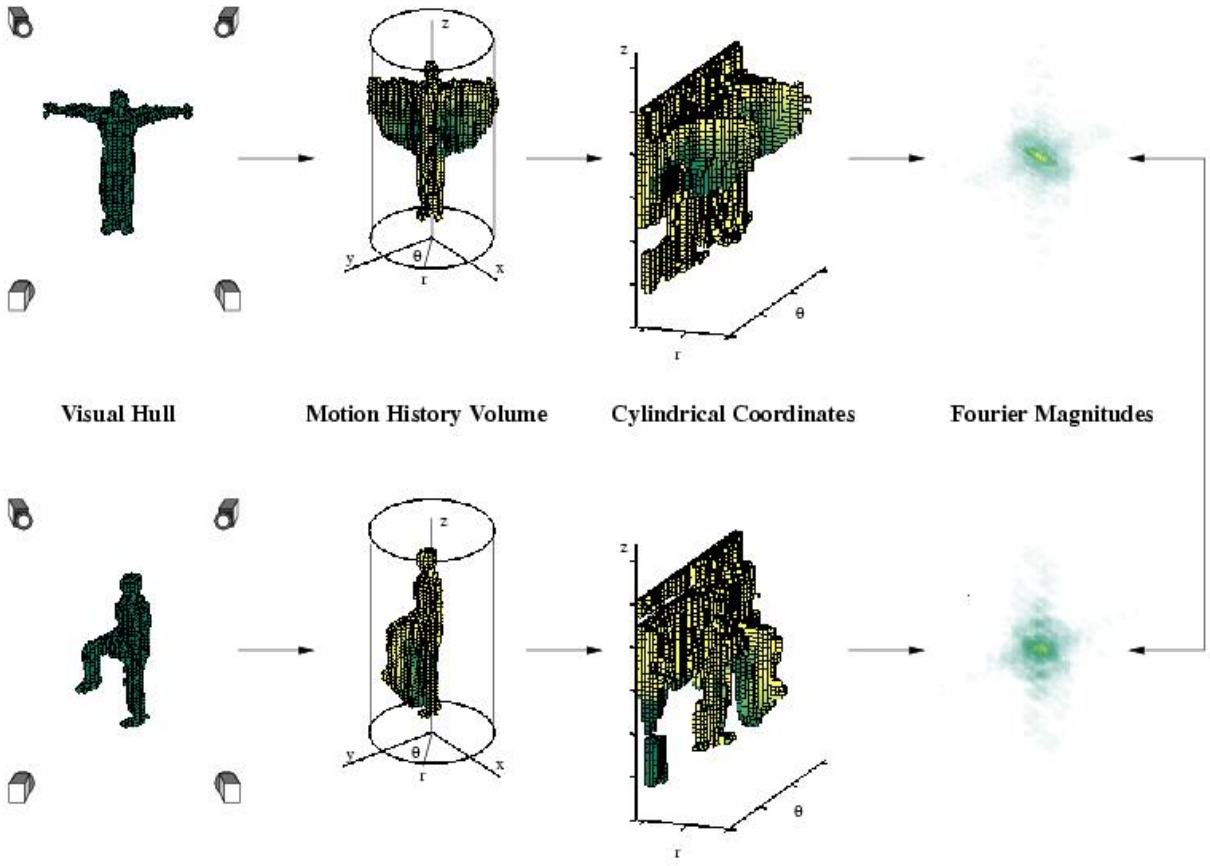

Figure 1: The two actions are recorded by multiple cameras, spatially integrated into their visual hulls, and temporally integrated into motion history volumes. Invariant motion descriptors in Fourier space are used for comparing the two actions.

egories in natural language [10].

The paper is organized as follows. First, we recall Davis and Bobick's definition of motion templates and extend it to three dimensions in Section 2. We present efficient descriptors for matching and aligning MHVs in Section 3. We discuss discrimination and recognition of basic human action classes in Section 4 with preliminary results which are discussed in Section 5.

2 Definitions

In this section, we first recall 2D motion templates as introduced by Bobick and Davis in [3] to describe temporal actions. We then propose their generalization to 3D in order to remove the viewpoint dependence in an optimal fashion using calibrated cameras.

2.1 Motion History Images

Motion Energy Images (MEI) and Motion-History Images (MHI) [3] were introduced to capture motion information in images. They encode, respectively, where motion occurred, and the history of motion occurrences, in the image. Pixel values are therefore binary values (MEI) encoding motion occurrence at a pixel, or multiple-values (MHI) encoding how recently motion occurred at a pixel. More formally, consider the binary-valued function $D(x, y, t)$ indicating motion at time t and location (x, y), then the MHI function is defined by:

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

where τ is the maximum duration a motion is stored. The associated MEI can easily be computed by thresholding $h > 0$.

The above motion templates are based on motion, i.e. $D(x, y, t)$ is a motion indicating function, however Bobick

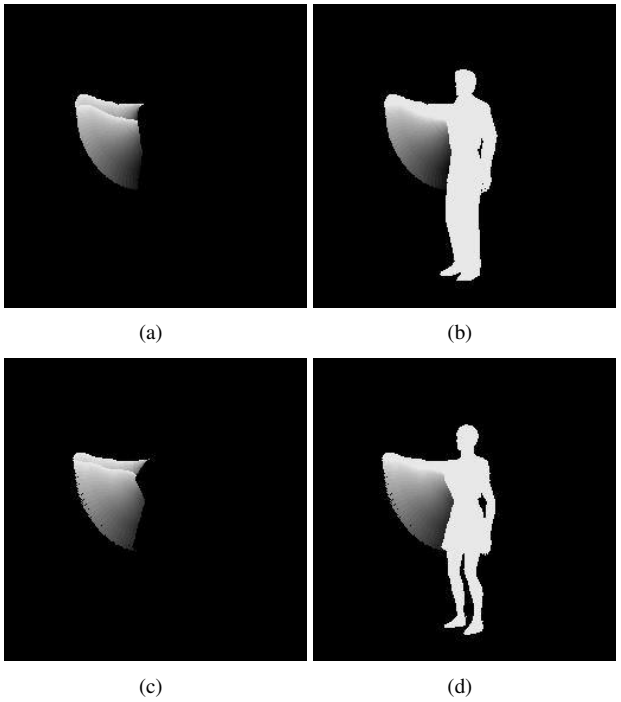

Figure 2: Motion versus occupancy. Using motion only in image (a), we can roughly gather that someone is lifting one arm. Using the whole silhouette instead, in (b), makes it clear that the right arm is lifted. However the same movement executed by a woman, in (c), compares favorably with the man's action in (a), whereas the whole bodies comparisons between (b) and (d) is less evident.

and Davis also suggest to compute templates based on occupancy, replacing $D(x, y, t)$ by the silhouette occupancy function. They argue that including the complete body makes templates more robust to incidental motions that occur during an action. Our experiments confirm that and show that occupancy provides robust cues for recognition, even if occupancy encodes not only motion but also shapes which may add difficulties when comparing movements, as illustrated in figure 2.

2.2 Motion History Volumes

In this paper, we propose to extend 2D motion templates to 3D. The choice of a 3D representation has several advantages over a single, or multiple, 2D view representation:

- A 3D representation is a natural way to fuse multiple images information. Such representation is more informative than simple sets of 2D images since additional calibration information is taken into account.
- A 3D representation is more robust to the object's positions relative to the cameras as it replaces a possibly complex matching between learned views and the actual observations by a 3D alignment (see next section).
- A 3D representation allows different camera configurations.

Motion templates extends easily to 3D by considering the occupancy function $D(x, y, z, t)$ in 3D and by considering voxels instead of pixels:

$$v_\tau(x, y, z, t) = \begin{cases} \tau & \text{if } D(x, y, z, t) \\ \max(0, h_\tau(x, y, z, t-1) - 1) & \text{otherwise.} \end{cases} \quad (2)$$

In the rest of the paper, we will assume templates to be normalized and segmented with respect to the duration of an action:

$$v(x, y, z) = v_{t_{\max} - t_{\min}}(x, y, z, t_{\max})/t_{\max}, \quad (3)$$

where $t_{\min}$ and $t_{\max}$ are start- and end time of an action.

The input occupancy function $D(x, y, z, t)$ is estimated using silhouettes and thus, corresponds to the visual hull [13]. Visual hulls present several advantages, they are easy to compute and they yield robust 3D representations. Note however that, as for 2D motion templates, different body proportions may still result in very different templates.

3 Motion Descriptors

Our objective is to compare body motions that are free in locations, orientations and sizes. This is not the case of motion templates, as defined in the previous section, since they encode space occupancy. The location and scale dependencies can be removed by centering and scale normalizing motion templates, as usual in shape matching. For the rotation, and following Bobick and Davis [3] who used the Hu Moments [8] as rotation invariant descriptors, we could consider their simple 3D extensions [17]. However, several works tend to show that moments are inappropriate feature descriptors, especially in the presence of noise, e.g. [19]. In contrast, Fourier based features have frequently demonstrated better results [5, 7]. Fourier based features are robust to noise and irregularities, and present the nice property to separate coarse global and fine local features in low and high frequency components. Moreover, they can be efficiently computed using fast Fourier-transforms (FFT). Our approach is therefore based on these features.

Recent works in shape matching [11] present feature vectors based on Fourier spherical harmonic representations. The idea is to voxelize centered and scale-normalized shapes into a spherical coordinate system. In such coordinate systems rotations simply map onto translations. By the fact 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 transformation:

$$F_t(k) = F_0(k)e^{-j2\pi kx_0}, \tag{4}$$

they take absolute values of Fourier spherical harmonics as rotation invariant descriptors.

In a similar way, we use Fourier-magnitudes and cylindrical coordinates, centered on bodies, to express motion templates in a way invariant to locations and rotations around the z-axis. The overall choice 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, but it appears to be reasonable in most situations. Furthermore, by restricting the Fourier-space representation to the lower frequencies, we also implicitly allow for additional degrees of freedom in object appearances and action executions. The following section details our implementation.

3.1 Invariant Representation

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

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

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

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

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.

From the 3D Fourier-transform $V(k_r, k_\theta, k_z)$ applied on the motion template v:

$$V(k_r, k_\theta, k_z) = \tag{5}$$
$$\int_{-\infty}^{\infty}\int_{-\pi}^{\pi}\int_{-\infty}^{\infty} v(r, \theta, z)e^{-j(k_r r + k_\theta \theta + k_z z)}drd\theta dz,$$

we take as invariant feature vector the magnitudes:

$$f_{k_r k_\theta k_z} = |V(k_r, k_\theta, k_z)|, \tag{6}$$

for some frequencies k_r, k_θ, k_z. Cylindrical projections and Fourier projections of MHVs are shown in figure 3. Examples on how to choose the frequencies $f_{k_r k_\theta k_z}$ are detailed in section 4.

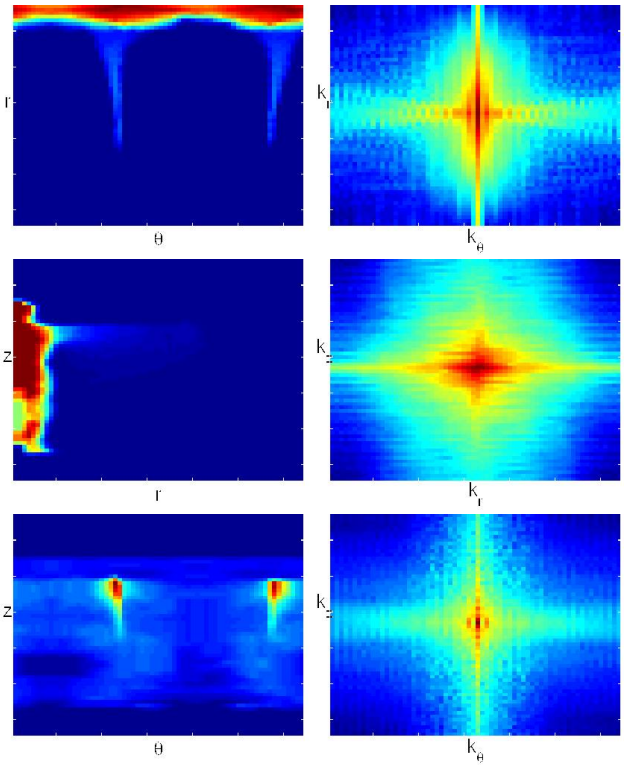

Figure 3: Volume and spectra of "lift both arms sideways" action: Cylindrical representation in $(\theta, r), (r, z), (\theta, z)$ averaged over the third dimension for visualization purposes (left column) and corresponding Fourier spectra (right column).

4 Classification Using Motion Descriptors

We have tested the presented descriptors and evaluated how discriminant they are with different actions, different bodies or different orientations. Results obtained on real data are presented in this section to give insights into the method's potential.

In the following experiments, an action data-set of two persons, performing 11 actions three times, was used. In order to have a high variance in the body sizes, the data-set was build with a small women and a tall man. Both where

Figure 4: Top: the two actors. Bottom: same action with different orientation.

asked to randomly choose different positions and orientations (within the camera's range) while repeating the actions (see table 1 and figures 4 and 5). No indications were given for speed, accuracy, exact start and end points, remaining body pose, etc.

	Actions
1	lift right arm ahead
2	lift right arm sideways
3	lift left arm sideways and ahead
4	lift left arm sideways
5	lift both arms ahead then sideways
6	drop 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

Table 1: Actions by number

Temporal segmentation of the sequences was done manually. Visual hulls were computed from 6 silhouettes, obtained using a standard background subtraction method. The resulting motion templates were mapped into a discrete cylindrical coordinate representation with size $64 \times 64 \times 64$. For reasons mentioned earlier (see Section 2.1), as well as due to flickering noise from background-subtraction, we only use motion history templates based on the whole body

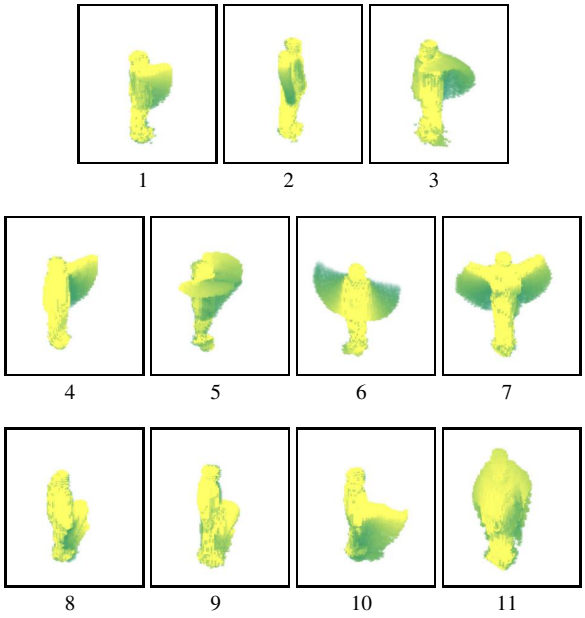

Figure 5: Perspective views of the motion history volumes computed for each action category.

silhouette. Binary MEVs were in particular excluded from the tests since they did not yield any improvement over MHVs.

4.1 Classification Using Euclidean Distances

We present first classification tests based on the complete normalized feature vector (6) of size $64 \times 64 \times 64 = 262144$, and applied on 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 by summing Euclidean distances over the feature vector's elements and with respect to the closest action class:

$$d(\mu, f) = \sum_i (\mu_i - f_i)^2, \qquad (7)$$

where μ represents a class mean and f a sample. Altogether the test returns 8 (3+5) false classifications, i.e. a classification rate of 87.9% (58 of 66). Figure 7 shows the distances and assignments in a confusion matrix. Figure 8 shows the averaged distances between classes and action executions. Interestingly, note how similar actions, or actions involving the same body parts, are grouped (dark areas).

In a second experiment, we reduced the size of the feature vectors to 216 elements representing the $6 \times 6 \times 6$ lowest frequencies. This empirical choice offers a good compro-

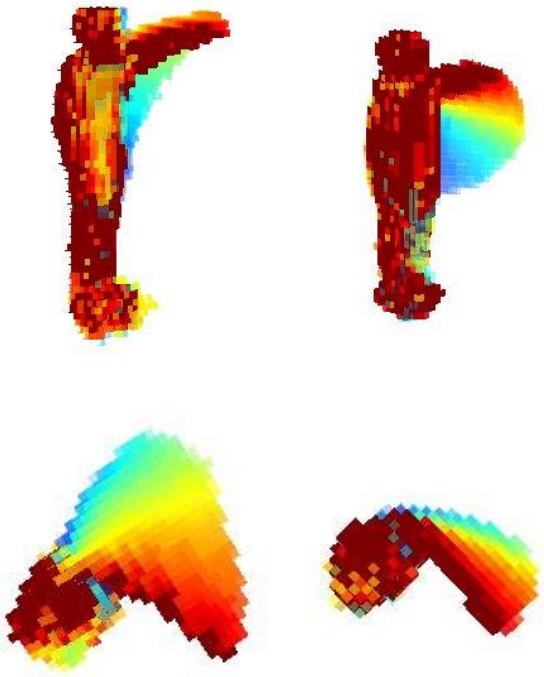

Figure 6: Differences in style between our two actors performing the same action (lift arm sideways and ahead). Top: side view. Bottom: top view. Left: male actor. Right: female actor.

mise between computational costs and classification rate: the obtained classification rate was 83.3%.

4.2 Classification Using Mahalanobis Distances

By taking the Euclidean distance, we implicitly assume that all vector elements weight equivalently. While true in an Euclidean space, such assumption may be wrong in the feature space we consider. Thus, experiments were also conducted with different weights associated to feature vector elements through the Mahalanobis distance. For each feature vector element/frequency, a global variance σ_i over the whole data-set is computed. Classification is then achieved using all frequencies having a variance higher than a small threshold ϵ and a simplified Mahalanobis-distance.

$$d(\mu, f) = \sum_{\sigma_i > \epsilon} \frac{(\mu_i - f_i)^2}{\sigma_i^2}. \quad (8)$$

The resulting vectors include around 800 features (man:803 woman:811). The classification rate is 90.9% in this case,

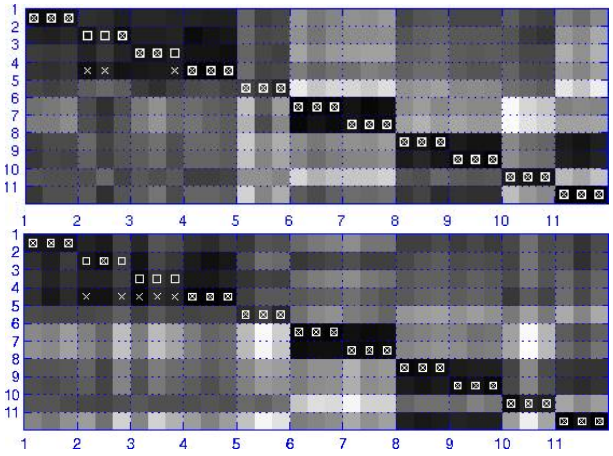

Figure 7: Top: Distances in feature space (Fourier magnitudes) between actions performed by female actor three times (horizontal axis) and categories learned from male actor (vertical axis). Ground-truth is indicated by squares, computed assignments are indicated by crosses. Bottom: Same figure for distances between "male" actions and "female" categories.

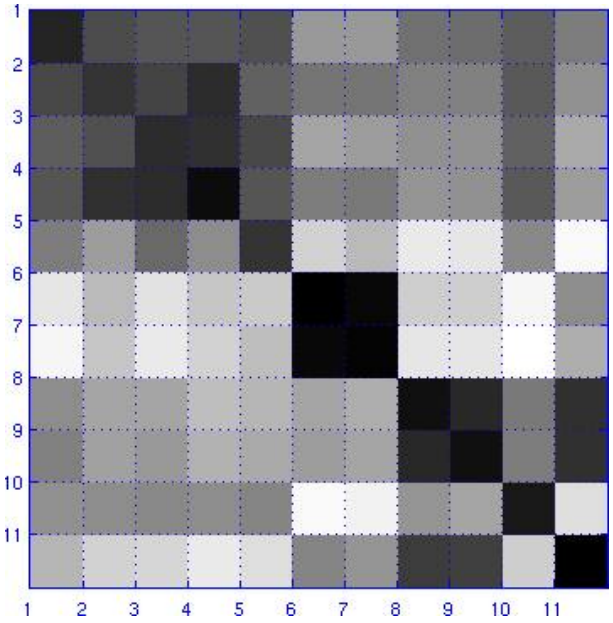

Figure 8: Average distances in feature space between action classes using all examples from figure 7.

with 6 remaining false assignments which mainly result from left-right arm confusions and high variances in action execution (see figure 6).

Finally, we also used 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 was computed:

$$\Sigma = \frac{1}{N}\sum^{N}(f-M)(f-M)^{\top}, \tag{9}$$

where M represents the mean value over all training samples. The Mahalanobis distance between the feature vector f and a class mean μ representing one action is:

$$d(\mu, f) = (f-\mu)^{\top} V \Lambda^{-1} V^{\top} (f-\mu),$$

where Λ contains 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.

Following this principle, we classified using 2 components only, getting 56.1% of right assignments. Using 12, the rate increased up to 84.8%, and the maximal rate of 93.9% was obtained with 16 components. Adding more components did not improve this results. Table 2 summarizes all results.

Method	Features	Classification Rate
euclidean distance	262144	87.9%
euclidean distance	216	83.3%
simplified Mahalanobis	800	90.9%
PCA and Mahalanobis	2	56.0%
PCA and Mahalanobis	12	84.8%
PCA and Mahalanobis	16 and up	93.9%

Table 2: Summary of Classification Results

5 Discussion and future work

With only two actors, a small vocabulary of actions and a few viewpoints, our experiments have presented us with a wealth of variations in appearance, which would be very difficult to confront using a single, uncalibrated camera. Our preliminary results indicate that motion history volumes are a useful representation for human action, and that invariant motion descriptors can be robustly extracted from them for comparing actions performed in different styles and recorded from different viewpoints. In future work, we plan to use those descriptors for learning statistical models of human actions from examples. This will require a substantial amount of work, both in the proper choice of the *primitive* action classes and in the optimal selection of features from the large set of Fourier descriptors used in this paper.

Although we cannot report results for lack of space, another important finding of our research is that viewpoint-invariant motion descriptors (Fourier magnitudes) are at least as efficient as correlation-based methods, at least for comparing simple actions. Numerous experiments have shown that, although it is possible to precisely recover the relative orientations between history volumes using phase or normalized correlation in Fourier space, and compare the aligned volumes directly, this almost never improves the classification results. Using invariant motion descriptors is of course advantageous because we do not need to align training examples for learning a class model, or align test examples with all class prototypes for recognition.

At the current stage of this research, our best results are obtained by discarding the phase information from the Fourier-transformed history volumes. Fourier magnitudes are a common choice in invariant image and shape matching, even though phase information usually contains more information and may actually better represent the differences between images and shapes [16]. Thus, an optimal descriptor should retain magnitude and *relative* phase information, ignoring only the absolute phase. Such an approach does not seem practical, because the ambiguity of the phase (the nth frequency component has an n-fold redundancy) makes computation of the relative phase a complex problem. Normalizing methods [1, 20] overcome this problem by shifting the phase of all objects, based on selected components, in a fixed position. However, initial experiments with phase dependent normalized descriptors didn't show significant improvements in classification results, and even seemed to have negative influence on the generalization process. In the light of those results, we believe a promising direction for classification of more complex actions will be to extract Fourier magnitudes in selected spatial locations, i.e. computing a low-resolution spatial spectrogram of the motion history volume. Interestingly, such an approach is consistent with recent findings in neuroscience [18] that object recognition in primates is mostly supported by Fourier magnitudes, not phase.

6 Conclusion

Using a limited but varied data set, we have been able to extract a small number of 3D motion descriptors that appear to support meaningful categorization of simple action classes performed by two actors, irrespective of viewpoint, gender and body sizes. Best results are obtained by discarding the phase in Fourier space and performing dimensionality reduction with PCA. The number of useful dimensions (12-16) is comparable to the number of action classes, and small enough that we can hope to train optimal classifiers for those actions from relatively small numbers of examples. In future work, we will study how this result scales

up, both in terms of the number of dimensions and in terms of classification rates, when the number of actions and performers is increased by several orders of magnitude. We anticipate that linear discriminant methods may prove useful in that respect.

References

[1] Y.S. Abu-Mostafa and D. Psaltis. Image normalization by complex moments. *PAMI*, 7(1):46–55, January 1985.

[2] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, 2001.

[3] James W. Davis. Hierarchical motion history images for recognizing human motion. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46, 2001.

[4] Alexei A. Efros, Alexander Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *ICCV*, 2003.

[5] A. E. Grace and M. Spann. A comparison between fourier-mellin descriptors and moment based features for invariant object recognition using neural networks. *Pattern Recognition Letters*, 12(10):635–643, 1991.

[6] Richard Green and Ling Guan. Quantifying and recognizing human movement patterns from monocular video images. *IEEE transaction on Circuits and Systems for Video Technology*, 14, february 2004.

[7] Daniel Heesch and Stefan M. Rueger. Combining features for content-based sketch retrieval - a comparative evaluation of retrieval performance. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 41–52, London, UK, 2002. Springer-Verlag.

[8] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8:179–187, February 1962.

[9] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), august 2000.

[10] R. Jackendoff. On beyond zebra: the relation of linguistic and visual information. *Cognition*, 20:89–114, 1987.

[11] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Symposium on Geometry Processing*, June 2003.

[12] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[13] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994.

[14] David Marr and Lucia Vaina. Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B*, 214:501–524, 1982.

[15] Jan Neumann, Cornelia Fermüller, and Yiannis Aloimonos. Animated heads: From 3d motion fields to action descriptions. In *DEFORM/AVATARS*, 2000.

[16] Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–41, 1981.

[17] F.A. Sadjadi and E.L. Hall. Three-dimensional moment invariants. *PAMI*, 2(2):127–136, 1980.

[18] L. Shams and C. von der Malsburg. The role of complex cells in object recognition. *Vision Research*, 42(22):2547–2554, 2002.

[19] Dinggang Shen and Horace Ho-Shing Ip. Discriminative wavelet shape descriptors for recognition of 2-d patterns. *Pattern Recognition*, 32(2):151–165, 1999.

[20] E.P. Simoncelli. A rotation invariant pattern signature. In *ICIP96*, pages 185–188, 1996.

[21] T.F. Syeda-Mahmood, M.A.O. Vasilescu, and S. Sethi. Recognition action events from multiple viewpoints. In *EventVideo01*, pages 64–72, 2001.

[22] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR05*, pages I: 984–989, 2005.