Free viewpoint action recognition using motion history volumes

Daniel Weinland, Rémi Ronfard, Edmond Boyer

Free Viewpoint Action Recognition using Motion History Volumes

Daniel Weinland [1], Remi Ronfard, Edmond Boyer

Perception-GRAVIR, INRIA Rhone-Alpes, 38334 Montbonnot Saint Martin, France.

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. Results indicate that this representation can be used to learn and recognize basic human action classes, independently of gender, body size and viewpoint.

Key words: action recognition, view invariance, volumetric reconstruction

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. According to Neumann et al. [1] and from a computational perspective, actions are best defined as four-dimensional patterns

Email addresses: weinland@inrialpes.fr (Daniel Weinland), ronfard@inrialpes.fr (Remi Ronfard), edmond.boyer@inrialpes.fr (Edmond Boyer).

[1] D. Weinland is supported by a grant from the European Community under the EST Marie-Curie Project Visitor.

in space and in time. 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 however 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.

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. We use multiple cameras and shape from silhouette techniques. We separate action recognition 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, as pointed out by Kojima et al. [2]. 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, e.g. stochastic parsing as proposed by Ivanov and Bobick [3]. In this paper, 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.

In previous work on motion descriptors, Green and Guan [4] use positions and velocities of human body parts, but such information is difficult to extract automatically during unrestricted human activities. Motion descriptors which can be extracted automatically, and which have been used for action recognition, are optical flows, as proposed by Efros et al. [5], motion templates in the seminal work of Bobick and Davis [6], and space-time volumes, introduced by Syeda-Mahmood et al. [7] or Yilmaz and Shah [8]. 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 [6], relative motion directions [5], and point correspondences [7,8]. 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

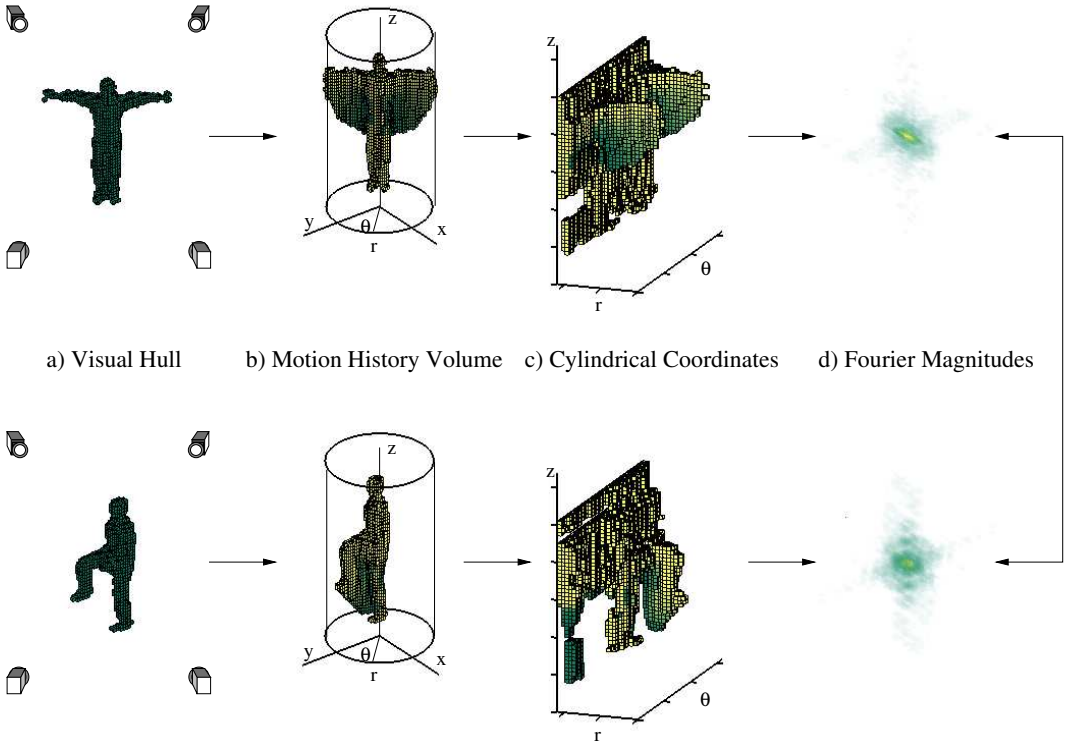


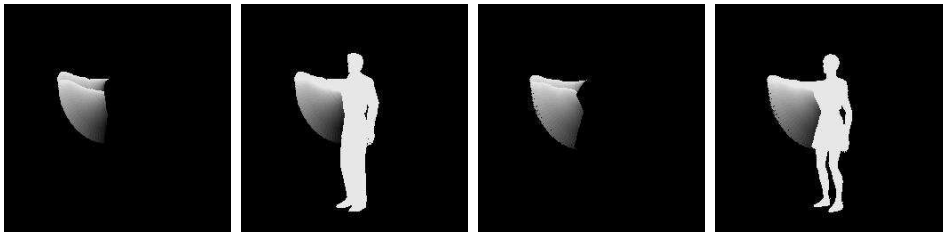
Fig. 1. The two actions are recorded by multiple cameras, spatially integrated into their visual hulls (a), and temporally integrated into motion history volumes (b)(c). Invariant motion descriptors in Fourier space (d) are used for comparing the two actions.

propose a representation based on Fourier analysis of motion history volumes in cylindrical coordinates. Figure 1 explains our method for comparing two action sequences. We separately compute their visual hulls and accumulate them into motion history volumes. We transform the MHVs into cylindrical coordinates around their vertical axes, and extract view-invariant features in Fourier space. Such a representation fits nicely within the framework of Marr's 3D model [9] which has been advocated by linguist Jackendoff [10] as a useful tool for representing action categories in natural language.

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 present classification results in Section 4 and conclude in Section 5.

2 Definitions

In this section, we first recall 2D motion templates as introduced by Bobick and Davis in [6] to describe temporal actions. We then propose their generalization



(a) (b) (c) (d)

Fig. 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.

to 3D in order to remove the viewpoint dependence in an optimal fashion using calibrated cameras. Finally, we show how to perform temporal segmentation using the 3D MHVs.

2.1 Motion History Images

Motion Energy Images (MEI) and Motion History Images (MHI) [6] 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 $D(x, y, t)$, $D = 1$ indicating motion at time t and location (x, y), then the MHI function is defined by:

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

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 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.

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, where $D = 1$ if (x, y, z) is occupied at time t and $D = 0$ otherwise, and by considering voxels instead of pixels:

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

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

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

where $t_{\min}$ and $t_{\max}$ are start and end time of an action. Hence, motions loose dependencies on absolute speed and result all in the same length. Section 2.3 shows how we detect these boundaries using a motion energy based segmentation.

The input occupancy function $D(x, y, z, t)$ is estimated using silhouettes and thus, corresponds to the visual hull [11]. 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. Figure 3 shows examples for motion history volumes.

2.3 Temporal Segmentation

Temporal Segmentation consist in splitting a continuous sequence of motions into elementary segments. In this work, we use an automatic procedure that we

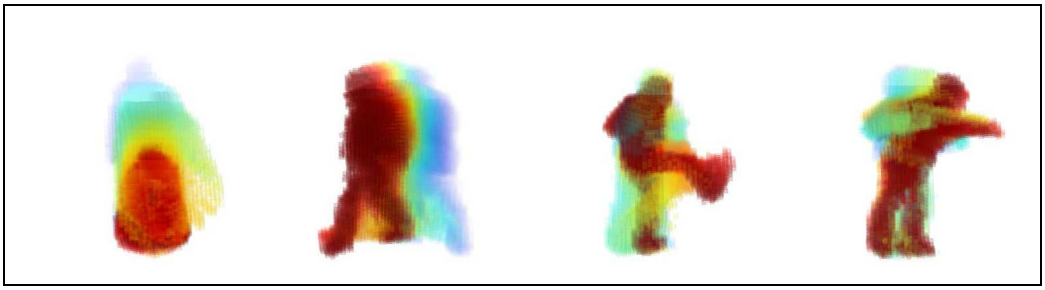


Fig. 3. Motion history volume examples: From left to right: "sit down"; "walk"; "kick"; "punch". Color values encode time of last occupancy.

recently introduced in [12]. It relies on the definition of motion boundaries as minima in motion energy, as originally proposed by Marr and Vaina [9]. Such minima correspond either to small rests between motions or to reversals in motion. As it turns out, an approximation of the global motion energy can be effectively computed using MHVs: Intuitively, instant motion can be encoded using MHVs over small time windows (typically 2 - 10 frames). Then the sum over all voxel values at time t will give a measure of the global motion energy at that time. Next, we search this energy for local minima, and recompute the MHVs based on the detected boundaries. For more details we refer to our work in [12].

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, with respect to the center of mass, and scale normalizing, with respect to a unit variance, motion templates, as usual in shape matching. For the rotation, and following Bobick and Davis [6] who used the Hu Moments [13] as rotation invariant descriptors, we could consider their simple 3D extensions by Sadjadi and Hall [14]. However, our experiments with these descriptors, based on first and second order moments, were unsuccessful in discriminating detailed actions. In addition, using higher order moments as in [15] is not easy in practice. Moreover, several works tend to show that moments are inappropriate feature descriptors, especially in the presence of noise, e.g. Shen [16]. In contrast, several works, such as that by Grace and Spann [17] and Heesch and Rueger [18], demonstrated better results using Fourier based features. 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.

Invariance of the Fourier transform follows from the Fourier *shift theorem*: 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:

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

Hence, Fourier magnitudes $|F_t(k)|$ are shift invariant signal representations. The invariance property translates easily onto rotation by choosing coordinate systems that map rotation onto translation. Popular example is the Fourier-Mellin transform, e.g. Chen et al. [19], that uses log-polar coordinates for translation, scale, and rotation invariant image registration. Recent work in shape matching by Kazhdan et al. [20] proposes magnitudes of Fourier spherical harmonics as rotation invariant shape 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:

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

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

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

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 $v(r, \theta, z) > 0$, 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 $\sigma_z = \sigma_r = 1$.

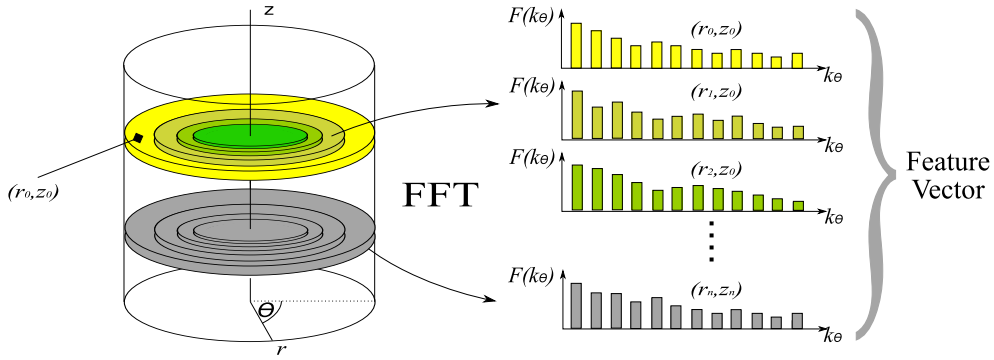


Fig. 4. 1D-Fourier transform in cylindrical coordinates. Fourier transforms over θ are computed for couples of values (r, z). Concatenation of the Fourier magnitudes for all r and z forms the final feature vector.

We choose to normalize in z and r direction, instead of a principal component 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.

The absolute values $|V(r, k_\theta, z)|$ of the 1D Fourier-transform

$$V(r, k_\theta, z) = \int_{-\pi}^{\pi} v(r, \theta, z) e^{-j2\pi k_\theta \theta} d\theta, \tag{5}$$

for each value of r and z, are invariant to rotation along θ.

See Figure 4 for an illustration of the 1D-Fourier transform. Note that various combinations of the Fourier transform could be used here. For the 1D Fourier-transform the spatial order along z and r remains unaffected. One could say, a maximum of information in these directions is preserved.

An important property of the 1D-Fourier magnitudes is its *trivial ambiguity* with respect to the reversal of the signal. Consequently, motions that are symmetric to the z-axis (e.g. move left arm - move right arm) result in the same motion descriptors. This can be considered either as a loss in information or as a useful feature halving the space of symmetric motions. However, our practical experience shows that most high level descriptions of human actions do not depend on this separation.

In cases where it is important to resolve left/right ambiguities a slightly different descriptor can be used. One such descriptor is the magnitude $|V(k_r, k_\theta, k_z)|$

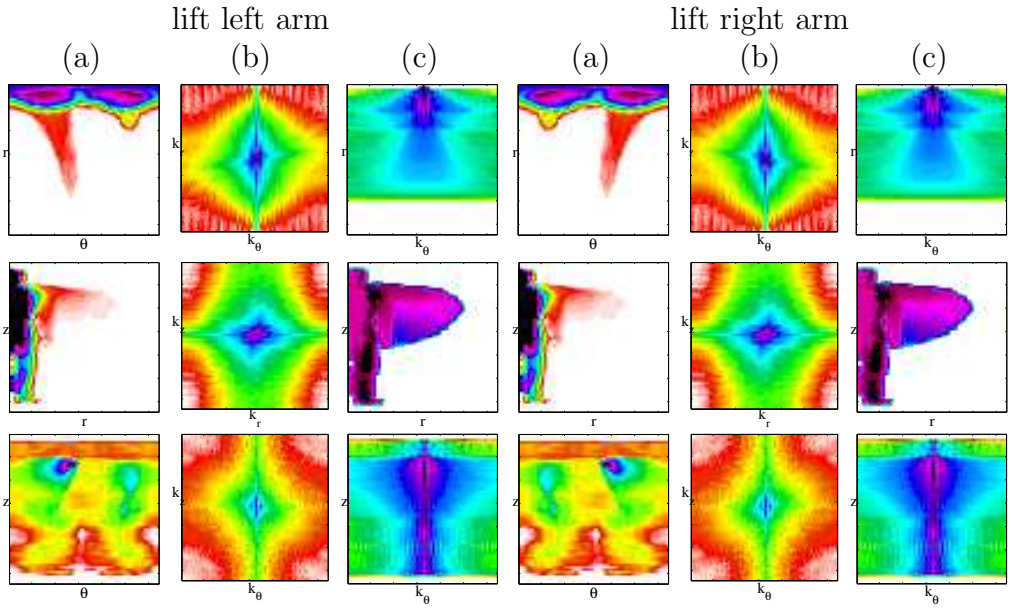


Fig. 5. Volume and spectra of sample motions: (a) cylindrical representation in $(\theta, r), (r, z), (\theta, z)$ averaged over the third dimension for visualization purposes; (b) corresponding 3D-Fourier Spectra; (c) 1D-Fourier spectra. Note that the 3D descriptor treats both motions differently (i.e. top and bottom row (b)), while the 1D descriptors treats them the same.

of the 3D-Fourier transform

$$V(k_r, k_\theta, k_z) = \int_{-\infty}^{\infty}\int_{-\pi}^{\pi}\int_{-\infty}^{\infty} v(r, \theta, z) e^{-j2\pi(k_r r + k_\theta \theta + k_z z)} dr d\theta dz, \tag{6}$$

applied to the motion template v. This descriptor is only symmetric with respect to an inversion of all variables, i.e. humans standing upside-down, which does not happen very often in practice. While our previous work [21] used that descriptor (6) with success, the results were anyway inferior to those obtained with (5) and an invariance to left right symmetry proved to be beneficial in many classification cases. A visualization of both descriptors is shown in Figure 5.

3.2 On Invariance vs. Exhaustive Search

Although we cannot report experiments for lack of space, another significant result of our research is that viewpoint-invariant motion descriptors (Fourier magnitudes) are at least as efficient as methods based on exhaustive search (correlation), 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 [22], 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.

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. Our previous results [21] using a small dataset of only two persons already indicated the high potential of the descriptor. This paper presents results on an extended dataset, the so called *IXMAS* dataset. The dataset is introduced in the next section, followed by classification results using dimensional reduction combined with Mahalanobis distance and linear discriminant analysis (LDA).

4.1 The IXMAS Dataset

The Inria Xmas Motion Acquisition Sequences (IXMAS) [2] aim to form a dataset comparable to the current "state-of-the-art" in action recognition. It contains 11 actions, see Figure 6 for instance, each performed 3 times by 10 actors (5 males / 5 females). To demonstrate the view-invariance, the actors freely change their orientation for each acquisition and no further indications on how to perform the actions beside the labels were given, as illustrated in Figure 7.

The acquisition was achieved using 5 standard Firewire cameras. Figure 8 shows example views from the camera setup used during the acquisition. From the video we extract silhouettes using a standard background subtraction technique modeling each pixel as a Gaussian in RGB space. Then visual hulls are carved from a discrete space of voxels, where we carve each voxel that not projects into all of the silhouettes. However, there are no special requirements for the visual hull computation and even the simplest method showed to work perfectly with our approach. After mapping into cylindrical coordinates the representation has a resolution of $64 \times 64 \times 64$. Temporal segmentation was performed as described in section 2.3. Note, that the temporal segmentations splits some of the actions into several elementary parts. To evaluate the descriptor on a selected dataset of primitive motions, we choose from each of the segments the one that best represents the motion. For example the action

[2] The data is available on the Perception website `http://perception.inrialpes.fr` in the "Data" section.

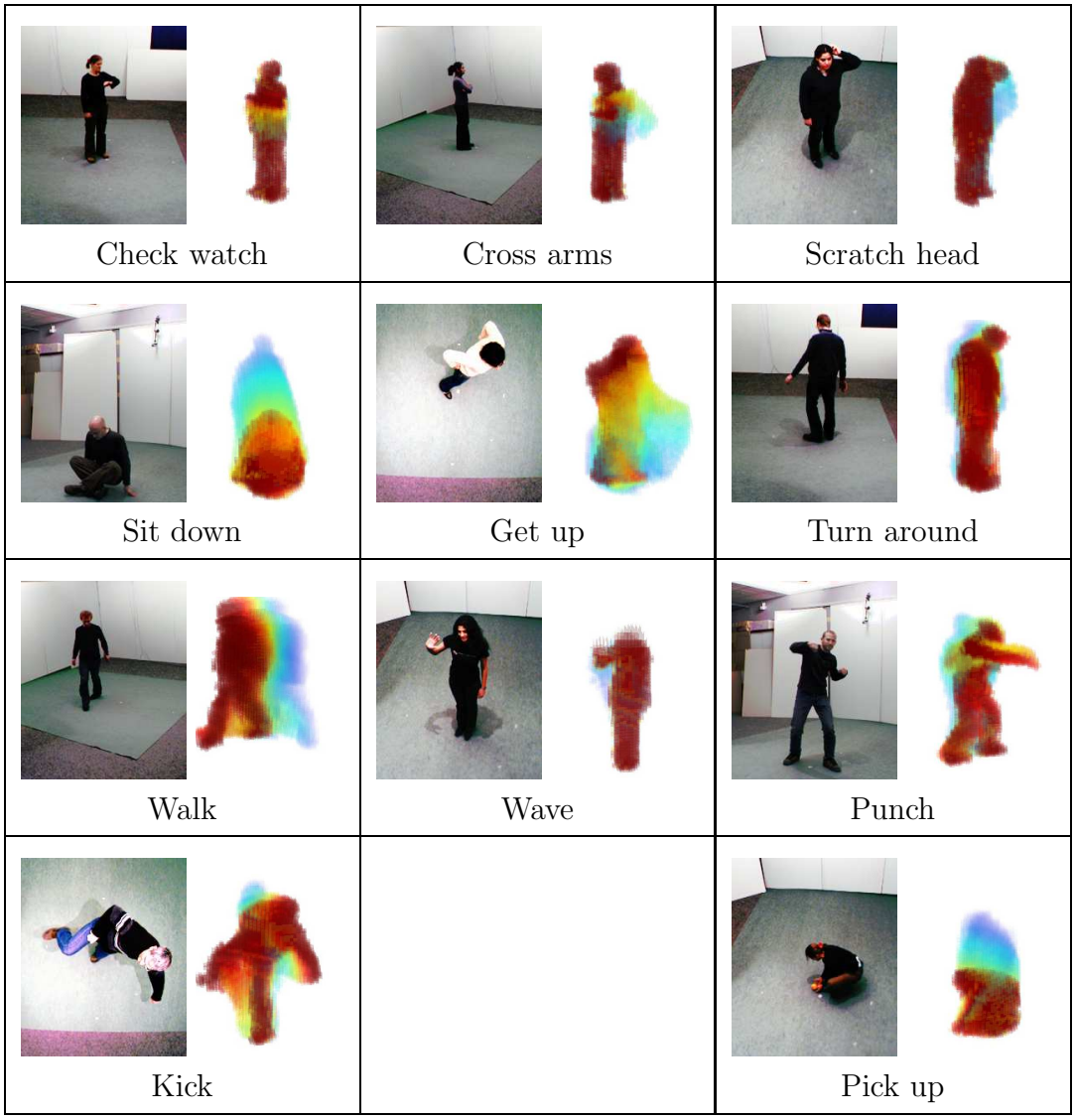


Fig. 6. 11 actions, performed by 10 actors.

"check watch" is split into three parts: an upward motion of the arm - several seconds of resting in this position - releasing the arm. From these motions we only use the first for the class "check watch". Another example is the action "walk", that has been broken down into separate steps. Interestingly, in those examples, we were able to classify even moderately complex actions based on one segment only. However, classification of composite actions is a topic of future research.

4.2 Classification Using Mahalanobis Distance and PCA

In initial experiments on a small dataset and with different distance measures (i.e. Euclidean distance, simplified Mahalanobis distance, and Mahalanobis

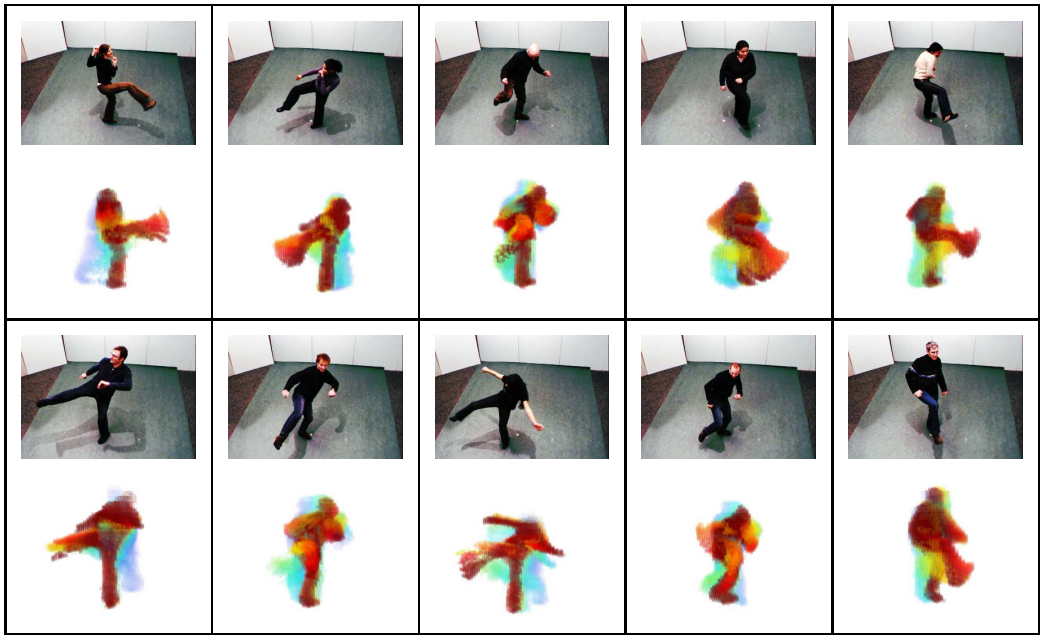


Fig. 7. Sample action "kick" performed by 10 actors.

Fig. 8. Example views of 5 cameras used during acquisition.

distance + PCA, see also [21]), the combination of a principal component analysis (PCA) dimensional reduction plus Mahalanobis distance based normalization showed best results. Due to the small amount of training samples we only used one pooled covariance matrix for all classes. Interestingly, we found that the method extends well to larger datasets and even competes with linear discriminant analysis (LDA), as will be shown in the next section.

PCA is a commonly used method for dimensional reduction. Data points are projected onto a subspace that is chosen to yield the reconstruction with minimum squared error. It has been shown that this subspace is spanned by the largest eigenvectors of the data's covariance Σ, and corresponds to the directions of maximum variance within the data. Further, by normalization with respect to the variance, an equally weighting of all components is achieved, similar to the classical use of Mahalanobis distances in classification, but here computed for one pooled covariance matrix.

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 according to a Mahalanobis distance associated to a PCA based dimensional reduction of the data vectors. One pooled covariance matrix Σ based on the training samples of all classes $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ was

#	Action	PCA	Mahal.	LDA
1	Check watch.	46.66%	86.66%	83.33%
2	Cross arms.	83.33%	100.00%	100.00%
3	Scratch head.	46.66%	93.33%	93.33%
4	Sit down.	93.33%	93.33%	93.33%
5	Get up.	83.33%	93.33%	90.00%
6	Turn around.	93.33%	96.66%	96.66%
7	Walk.	100.00%	100.00%	100.00%
8	Wave hand.	53.33%	80.00%	90.00%
9	Punch.	53.33%	96.66%	93.33%
10	Kick.	83.33%	96.66%	93.33%
11	Pick up.	66.66%	90.00%	83.33%
average rate		73.03%	93.33%	92.42%

Table 1
IXMAS data classification results. Results on PCA, PCA + Mahalanobis distance based normalization using one pooled covariance, and LDA are presented.

computed:

$$\Sigma = \frac{1}{n}\sum_{i}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top, \tag{7}$$

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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, $k \leq n-1$, and V the corresponding eigenvectors of Σ. Thus feature vectors are reduced to k principal components.

Following this principle, and reducing the initial descriptor (equation (5)) to $k = 329$ components an average classification rate of 93.33% was obtained with leave-one-out cross validation, where we successively used 9 of the actors to learn the motions and the 10th for testing. Note that in the original input space, as well as for a simple PCA reduction without covariance normalization the average rate is only 73.03%. Detailed results are given in Table 1.

For further data reduction, class specific knowledge becomes important in learning low dimensional representations. Instead of relying on the eigen-decomposition of one pooled covariance matrix, we use here a combination of PCA and Fisher linear discriminant analysis (LDA), see e.g. Swets and Weng [23], for automatic feature selection from high dimensional data.

First PCA is applied, $Y = V^\top X$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$, to derive a $m \leq n - c$ dimensional representation of the data points x_i, $i = 1, \ldots, n$. The class-number c dependent limit is necessary to guaranty non-singularity of matrices in discriminant analysis.

Fisher discriminant analysis defines as within-scatter matrix:

$$S_w = \sum_i^c \sum_j^{n_i} (\mathbf{y}_j - \mathbf{m}_i)(\mathbf{y}_j - \mathbf{m}_i)^\top, \tag{8}$$

and between-scatter matrix:

$$S_b = \sum_i^c (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top, \tag{9}$$

and aims at maximizing the between-scatter while minimizing the within-scatter, i.e. we search a projection W that maximize $\frac{\det(S_b)}{\det(S_W)}$. It has been proven that W equal to the largest eigenvectors of $S_w^{-1} S_b$ maximizes this ratio. Consequently a second projection $Z = W^\top Y$, $W = [w_1, \ldots, w_k]$, $k \leq c - 1$ is applied to derive our final feature representation Z.

During classification each class is represented by its mean vector $\mathbf{m}_i$. Any new action $\mathbf{z}$ is then classified by summing Euclidean distances over the discriminant features and with respect to the closest action class:

$$d(\mathbf{m}_i, \mathbf{z}) = ||\mathbf{m}_i - \mathbf{z}||^2. \tag{10}$$

In the experiments the magnitudes of the Fourier representation (equation (5)) are projected onto $k = 10$ discriminant features. Successively we use 9 of the actors to learn the motions, the 10th is used for testing. The average rate of correct classifications is then 92.42%. Class specific results are shown in Table 1 and Figure 9.

We note that we obtain much better results with the Mahalanobis distance, using the 329 largest components of the PCA decomposition, as compared to using the PCA components alone. LDA allows us to further reduce the

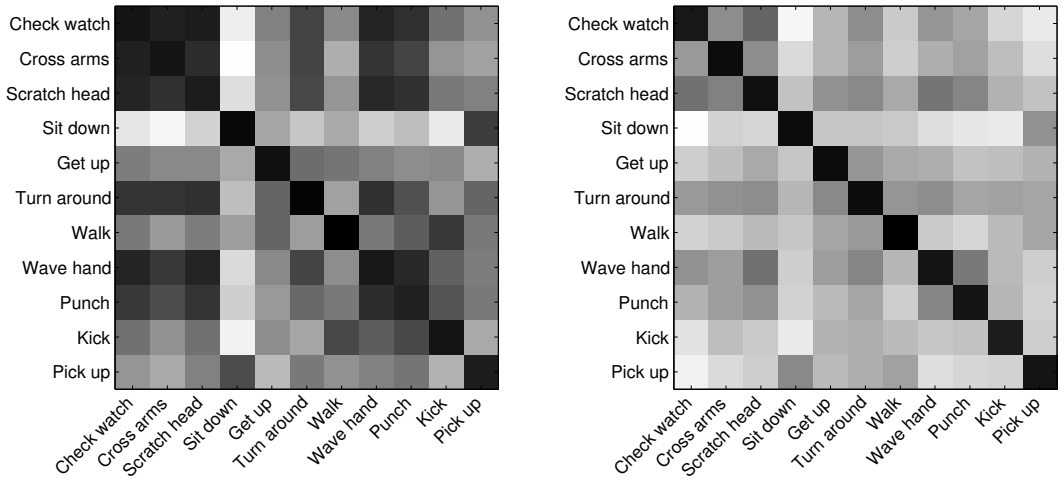


Fig. 9. Average class distance: (Left) before discriminant analysis. (Right) after discriminant analysis.

number of features to 10, but otherwise does not further improve the overall classification results.

4.4 Motion History vs. Motion Energy and Key Frames

With the same dataset as before, we compare our MHV based descriptors with a combination of key poses and energy volumes. While Davis and Bobick suggested in the original paper the use of history and binary images, our experiments with motion volumes showed no improvement in using a combination of MHVs and the binary MEVs. We repeated the experiment described in section 4.3, for MEVs. Using the binary information the recognition rate becomes 80.00% only. See Table 2 for detailed results. As can be expected: reverse actions, e.g. "sit down" - "get up", present lower scores with MEVs than with MHVs. The MHVs show also better performance in discriminating actions on more detailed scales, e.g. "scratch head" - "wave".

Also, to show that integration over time plays a fundamental role of information, we compare our descriptor with descriptors based on a single selected *key frame*. The idea of key frames is to represent a motion by one specific frame, see e.g. Carlson and Sullivan [24]. As invariant representation, we use the magnitudes of equation (5). For the purpose of this comparison we simply choose the last frame of each MHV computation as corresponding *key frame*. The average recognition rate becomes 80.30%. While motion intensive action, e.g. "walk" - "turn around" score much lower, a few pose expressive actions, e.g. "pick up", achieve a better score. This may indicate that not all actions should be described with the same features.

#	Action	MEV	Key frame	MHV
1	Check watch.	86.66%	73.33%	86.66%
2	Cross arms.	80.00%	93.33%	100.00%
3	Scratch head.	73.33%	86.66%	93.33%
4	Sit down.	70.00%	93.33%	93.33%
5	Get up.	46.66%	53.33%	93.33%
6	Turn around.	90.00%	60.00%	96.66%
7	Walk.	100.00%	80.00%	100.00%
8	Wave hand.	80.00%	76.66%	80.00%
9	Punch.	93.33%	80.00%	96.66%
10	Kick.	90.00%	90.00%	96.66%
11	Pick up.	70.00%	96.66%	90.00%
average rate		80.00%	80.30%	93.33%

Table 2
IXMAS data classification results. Results using the proposed MHVs are presented. For comparison we also include results using binary MEVs and key frame descriptors.

We conclude, that invariant Fourier descriptors of binary motion volumes and key frames are suitable for motion recognition as well. However, the use of additional motion information, as present in the motion history volumes, in both cases distinctly improves the recognition.

4.5 Classification on Video Sequences

The previous experiments show that the descriptor performs well in discriminating selected sets of learned motions. In this experiment we test the descriptor on unseen motion categories as they appear in realistic situations. For this purpose we work on the raw video sequences of the IXMAS dataset. In a first step the dataset is segmented into small motion primitives using the automatic segmentation. Then each segment is either recognized as one of the 11 learned classes or rejected. As in the previous experiments, we work in PCA space spanned by the 11 sample motions and perform nearest-mean assignment. To decide for the "reject"-class we use a global threshold on the distance to the closest class.

The automatic segmentation of the videos results in 1188 MHVs, corresponding to approximately 23 minutes of video. In manual ground truth labeling

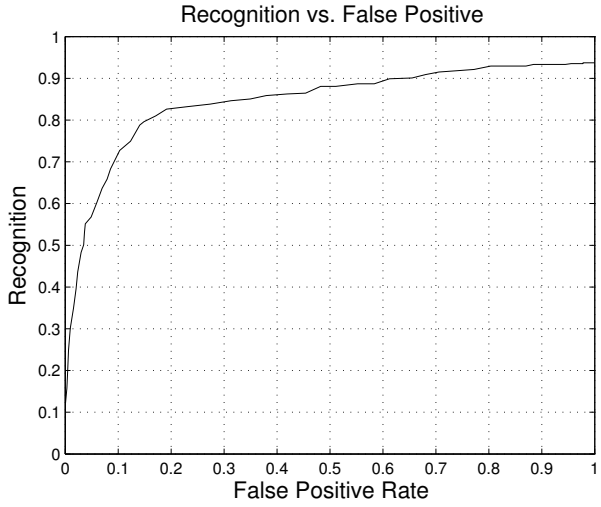


Fig. 10. Recognition on raw video sequences: Plots recognition rate into 11 classes against false positive rate.

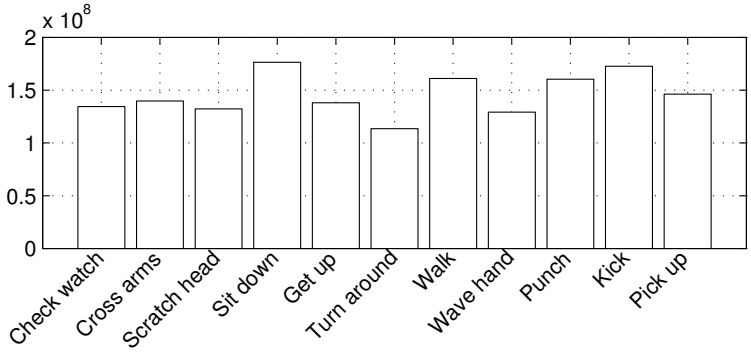


Fig. 11. Average distance between "reject"-samples and training classes.

we discover 495 known motions and 693 "reject"-motions. Note, that such a ground truth labeling is not always obvious. A good example is the "turn"-motion that was included in the experiments, but additional turn-like motions also appear as the actors where free to change position during the experiments. Moreover, it might be that an actor was accidentally checking his watch or scratching his head.

Testing in a leave-one-out manner, using all possible combinations of 9 actors for training and the remaining 10th for testing, we show a multi-class ROC curve, Figure 10, plotting the average number of correctly classified samples, against the number of false positives. We found a maximal overall recognition rate (including correctly rejected motions) of 82.79%, for 14.08% false positives and 78.79% correctly classified motions. Figure 11 shows the average distance between the "reject"-motions and the learned classes.

The experiments demonstrate the ability of MHVs even to work with large amounts of data and under realistic situations (23 minutes of video, 1188

motion descriptors). The segmentation proved to almost always detect the important parts of motions; MHVs showed good quality in discriminating learned and unseen motions.

An obvious problem for the false detections, is the nearly infinite class of possible motions. Modeling unknown motions may require more than a single threshold and class, multiple classes and explicit learning on samples of unknown motions becomes important. Another problem we found is, that many motions can not be modeled by a single template. Small motions may seem very similar, but over time belong to very different actions. For example the turn around motion is split into several small steps that may easily be confused with a single side step. In such cases temporal networks over templates, as e.g. in an HMM approach, must be used to resolve these ambiguities. However, we leave this for future work.

5 Conclusion

Using a data set of 11 actions, we have been able to extract 3D motion descriptors that appear to support meaningful categorization of simple action classes performed by different actors, irrespective of viewpoint, gender and body sizes. Best results are obtained by discarding the phase in Fourier space and performing dimensionality reduction with a combination of PCA and LDA. Further, LDA allows a drastic dimension reduction (10 components). This suggests that our motion descriptor may be a useful presentation for view invariant recognition of an even larger class of primitive actions. Our current work is suited to segmentation of composite actions into primitives, and classification of sequences of the corresponding LDA coefficients.

References

[1] J. Neumann, C. Fermller, Y. Aloimonos, Animated heads: From 3d motion fields to action descriptions, in: DEFORM/AVATARS, 2000.

[2] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions., International Journal of Computer Vision 50 (2) (2002) 171–184.

[3] Y. A. Ivanov, A. F. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 852–872.

[4] R. D. Green, L. Guan, Quantifying and recognizing human movement patterns from monocular video images-part i: a new framework for modeling human

motion., IEEE Trans. Circuits Syst. Video Techn. 14 (2) (2004) 179–190.

[5] A. A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: IEEE International Conference on Computer Vision, 2003.

[6] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

[7] T. Syeda-Mahmood, M. Vasilescu, S. Sethi, Recognition action events from multiple viewpoints, in: EventVideo01, 2001, pp. 64–72.

[8] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. I: 984–989.

[9] D. Marr, L. Vaina, Representation and recognition of the movements of shapes, Proceedings of the Royal Society of London B 214 (1982) 501–524.

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

[11] A. Laurentini, The Visual Hull Concept for silhouette-based Image Understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (2) (1994) 150–162.

[12] D. Weinland, R. Ronfard, E. Boyer, Automatic discovery of action taxonomies from multiple views, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006.
URL `http://perception.inrialpes.fr/Publications/2006/WRB06`

[13] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory IT-8 (1962) 179–187.

[14] F. Sadjadi, E. Hall, Three-dimensional moment invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (2) (1980) 127–136.

[15] C. Lo, H. Don, 3-d moment forms: Their construction and application to object identification and positioning, PAMI 11 (10) (1989) 1053–1064.

[16] D. Shen, H. H.-S. Ip, Discriminative wavelet shape descriptors for recognition of 2-d patterns., Pattern Recognition 32 (2) (1999) 151–165.

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

[18] D. Heesch, S. 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, Springer-Verlag, London, UK, 2002, pp. 41–52.

[19] Q. Chen, M. Defrise, F. Deconinck, Symmetric phase-only matched filtering of fourier-mellin transforms for image registration and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 16 (12) (1994) 1156–1168.

[20] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz, Rotation invariant spherical harmonic representation of 3d shape descriptors, in: Symposium on Geometry Processing, 2003.

[21] D. Weinland, R. Ronfard, E. Boyer, Motion history volumes for free viewpoint action recognition, in: IEEE International Workshop on modeling People and Human Interaction, 2005.
URL `http://perception.inrialpes.fr/Publications/2005/WRB05`

[22] C. D. Kuglin, D. C. Hines, The phase correlation image alignment method, in: IEEE International Conference on Cybernetics and Society, 1975, pp. 163–165.

[23] D. L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 831–836.

[24] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, in: Workshop on Models versus Exemplars in Computer Vision, 2001.