ANALYZING INVARIANCE OF FREQUENCY DOMAIN BASED FEATURES FROM VIDEOS WITH REPEATING MOTION

Kahraman Ayyildiz and Stefan Conrad
Department of Databases and Information Systems
Heinrich Heine University Duesseldorf, Universitätsstraße 1, 40225 Duesseldorf, Germany

Keywords: Motion Detection, Motion Recognition, Action Recognition, Repeating Movement, Video Classification, Frequency Feature, Invariance, View-invariant.

Abstract: This paper discusses an approach, which allows classifying videos by frequency spectra. Many videos contain activities with repeating movements. Sports videos, home improvement videos, or videos showing mechanical motion are some example areas. Motion of these areas usually repeats with a certain main frequency and several side frequencies. Transforming repeating motion to its frequency domain via FFT reveals these frequency features. In this paper we explain how to compute frequency features for video clips and how to use them for classifying. The experimental stage of this work focuses on the invariance of these features with respect to rotation, reflection, scaling, translation and time shift.

1 INTRODUCTION

Computer vision is a highly investigated research area in computer science. Some aspects of this area are video retrieval, video surveillance, human-computer interfaces, object tracking and action recognition. All of these topics play an important role for industry and technique. Video databases for instance can be found in major corporations or online video portals. Moreover video surveillance is needed to protect company buildings, public places and private properties. Today most digital video cameras use face tracking methods in order to focus on faces and zoom into important picture parts. So computer vision has relevance for public and private life.

In this work we explain an approach, which is able to detect, track and classify motion in video sequences. It is based on our previous research work (Ayyildiz and Conrad, 2011) and improves its feature extraction methods. Instead of considering single frequency maxima now we utilize complete frequency spectra derived from motion. Thus accuracies can be improved and the system is more robust against different types of invariance. Our approach focuses on repeating motion and resulting frequency features. It works for every motion type and is not limited to human gait recognition as described in (Meng et al., 2006; Zhang et al., 2004). The experimental part of our research work analyzes invariance aspects of these frequency features in order to find out, how robust the method works with varying camera settings. This aspect is important, since video databases ordinarily contain videos with different camera angles, zooming factors or object positions.

As a first step our method detects regions with motion for each frame. This regions lead to image moments for each frame, where a series of image moments represents a function. By a fast Fourier transform (FFT) this function is transformed to its frequency domain. A partitioning of this frequency domain into intervals gives different average amplitudes for each interval. These average amplitudes are considered as features for each clip. Once feature vectors are determined, a classifier can assign each clip to a class.

In the following section we focus on the whole process of video classification. We explain feature extraction phase and classification phase stepwise. Furthermore we define image moments and so-called *1D-functions* for transformation in section 3. The basic feature vectors *AAFIs* (Average Amplitudes of Frequency Intervals) are explained in section 4. Afterwards we introduce our radius based classifier *RBC* in section 5. The evaluation of our approach takes place in section 6. The following section discusses work related to our approach, where the last section reviews the presented methods.

Ayyildiz K. and Conrad S..
ANALYZING INVARIANCE OF FREQUENCY DOMAIN BASED FEATURES FROM VIDEOS WITH REPEATING MOTION.
DOI: 10.5220/0003798806590666
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 659-666
ISBN: 978-989-8565-03-7

2 CLASSIFYING VIDEOS BY AAFIS

In this section we explain methods used for our approach, where fig. 1 offers an overview of the different stages.

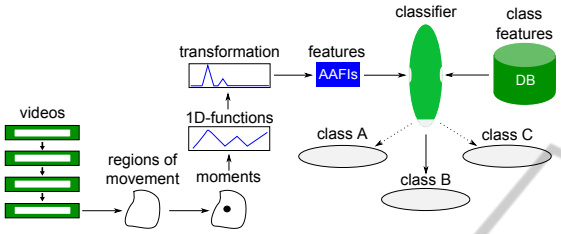

Figure 1: Flow diagram of whole classification process.

The goal of the whole classification process is to classify video sequences with repeating movements properly. Some examples for activity with repeating movement are jumping, playing tennis or hammering. First regions of movement are detected in every clip for each frame. Regions are detected by measuring the color differences of pixels in two frames following each other (see section 3.1). Based on these regions we calculate image moments, where two types of moments are applied: centroids and pixel variances. A chronological series of these moments are considered as 1D-functions and represent the motion in a video sequence. The FFT of one 1D-function reveals its frequency domain. By partitioning the frequency axis into intervals of same length, average amplitudes for each interval are computed. We name these averages *AAFIs* (Average Amplitudes of Frequency Intervals). AAFIs constitute the final feature vectors for each clip with respect to its motion. After determining the feature vector of a video its next class is computed by a classifier.

3 IMAGE MOMENTS AND 1D-FUNCTIONS

Frequency spectra result from repeating motion in video scenes and this motion has to be detected frame by frame at first. Once the motion is localized image moments and resulting 1D-functions can be figured. Next we define regions of motion and explain how these regions lead to 1D-functions.

3.1 Regions of Motion

Fig. 2 shows a person troweling a wall in two consecutive frames. By analyzing these two frames we detect regions with motion. Color differences between the first and the second frame are measured for each pixel. If the color difference of a pixel exceeds a predefined threshold and if there are enough neighbor pixels with a color difference beyond the same threshold, this pixel is considered to be a part of a movement. Thus a region of motion is represented by the conflation of pixels with motion.

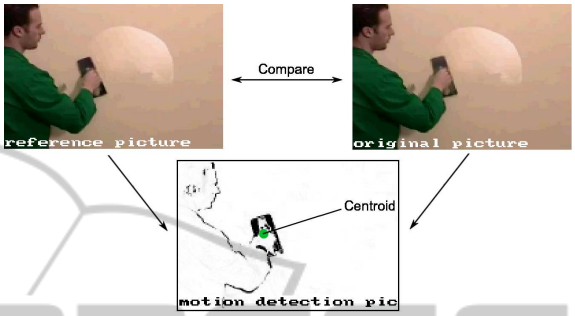

Figure 2: Regions with pixel activity and centroid.

Comparing the two frames results in a binary image arising from regions with movement. Moreover the centroid of regions with motion lies exactly on the right hand, because the most active areas are the arm, the hand, and the trowel. Hence the troweling determines the motion path of the centroid.

3.2 Image Moments

In image processing an image moment is the weighted average of picture pixel values. It is used to describe the area, the bias, or the centroid of segmented image parts. We distinguish two types of image moments: raw moments and central moments. Raw moments are sensitive to translation, whereas central moments are translation invariant. Next equation defines a raw moment M_{ij} for a two dimensional binary image $b(x,y)$ and $i, j \in \mathbb{N}$ (Wong et al., 1995):

$$M_{ij} = \sum_x \sum_y x^i \cdot y^j \cdot b(x,y) \qquad (1)$$

The order of M_{ij} is always $(i+j)$. M_{00} determines the area of segmented parts. Hence $(\overline{x},\overline{y}) = (M_{10}/M_{00}, M_{01}/M_{00})$ defines the centroid of segmented parts. Moreover the computation of central moments applies centroid coordinates (Wong et al., 1995).

$$\mu_{ij} = \sum_x \sum_y (x-\overline{x})^i \cdot (y-\overline{y})^j \cdot b(x,y) \qquad (2)$$

Here μ_{20} and μ_{02} represent the variances of pixels with regard to x and y coordinates, respectively.

3.3 Deriving 1D-functions

Function f is called a 1D-function, if it represents a

series of one-dimensional moment values. This series corresponds to the chronological order of frames in a video, which leads to function $f(t)$ with t as time. For $(\bar{x}_t, \bar{y}_t) = (M_{10_t}/M_{00_t}, M_{01_t}/M_{00_t})$ as the centroid coordinates depending on time t function $f_c(t) = (\bar{x}_t, \bar{y}_t)$ can be decomposed as follows:

$$f_{c_x}(t) = \bar{x}_t \wedge f_{c_y}(t) = \bar{y}_t \tag{3}$$

For the experimental stage in section 6 we use $f_{c_x}(t)$ and $f_{c_y}(t)$ instead of $f_c(t)$, because the transform of 1D-functions results in more decisive frequency spectra than transforming 2D-functions. For the same reason two separate 1D-functions of central moments are implemented and tested:

$$f_{v_x}(t) = \mu_{20_t}, \; f_{v_y}(t) = \mu_{02_t} \tag{4}$$

For any 1D-function $f(t)$ the direction of a moment at time t is defined by 5.

$$f_d(t) = \begin{cases} +1, & \texttt{if } f(t) - f(t-1) > 0 \\ 0, & \texttt{if } f(t) - f(t-1) = 0 \\ -1, & \texttt{if } f(t) - f(t-1) < 0 \end{cases} \tag{5}$$

4 AAFIS AS FEATURE VECTORS

This section explains how we compute feature vectors for videos by 1D-functions. As mentioned before the transform of a 1D-function via FFT spans a frequency spectrum. By partitioning this spectrum into intervals with same length, an average amplitude for each interval can be stated.

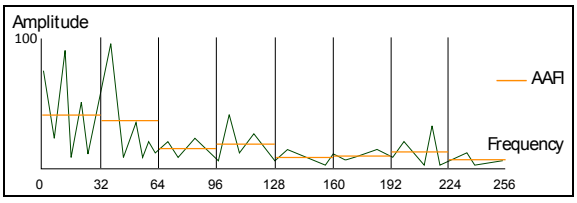

Figure 3: Average amplitudes of frequency intervals (AAFIs).

Fig. 3 illustrates this idea by dividing a frequency spectrum with a length of $m = 256$ units into $n = 8$ intervals. As we use the FFT variables m and n have to be a power of 2, where $m \geq n$. Moreover each orange line marks one average amplitude of one interval. This average amplitude is called AAFI (Average Amplitude of Frequency Interval). Thus with respect to our illustration in fig. 3 one 1D-function results in 8 average amplitudes respectively in one 8-dimensional feature vector. As mentioned in section 3 each video is described by two 1D-functions, the first one relates to the x-axis motion and the second one to the y-axis motion. So two 8-dimensional feature vectors can be stated, which results in a combined 16-dimensional feature vector for this example. It can be generalized that the partitioning of any frequency spectrum into n intervals leads to a $(2 \cdot n)$-dimensional feature vector for each video.

In our previous work (Ayyildiz and Conrad, 2011) we used up to 6 frequency maxima for each video as feature vector. Now the whole frequency spectrum is described by AAFIs and feature vectors reveal much more information about the motion type.

5 RADIUS BASED CLASSIFIER

Now we introduce our *Radius Based Classifier* (RBC). During the experimental phase this classifier turned out as very effective. The classifier measures the density of objects inside a predefined radius around an object, which has to be classified. This density is used for distance calculations.

5.1 Idea

Fig. 4 illustrates how the RBC works: So as to classify an object $o_a \in B$ the RBC assigns it to each existing class C_i. These assignments give rise to distances $dist(o_a, C_i)$. The more objects are located within radius ε, the smaller the distance.

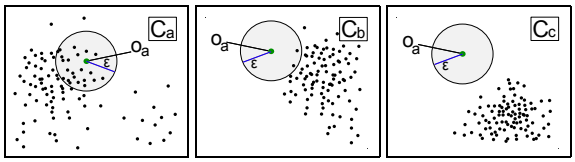

Figure 4: Classifying with RBC.

In fig. 4 there are three different example classes C_a, C_b, $C_c \in C$, where each class has its own typical object distribution. Assigning o_a to class C_a reveals, that there are many objects within the radius. In C_b the metric encloses just 2 objects. In class C_c objects are far away from o_a, so there is no object of this class within the Euclidian metric. According to these three classes, o_a fits best into class C_a, because it is part of the typical object distribution. At the same time this fact leads to the smallest distance.

5.2 Formalization

First we define $C = \{C_1, \ldots, C_m\}$ as our set of classes. Each class $C_i \in C$ contains a set of objects, so we define $C_i = \{o_{i_1}, \ldots, o_{i_{n_i}}\}$, $C_i \neq \{\}$ and $C_i \cap C_j = \{\}$ for $i \neq j$. The total of all objects in classes constitutes

our training set $A = C_1 \cup \ldots \cup C_m$. Test set objects in $B = \{o_1, \ldots, o_l\} \neq \{\}$ with $A \cap B = \{\}$ do not belong to any class.
Let object $o_b \in B$ and class $C_i \in C$, then radius ε determines the ε-neighborhood $N_\varepsilon(o_b, C_i)$. This ε-neighborhood encloses all objects of class C_i inside the predefined radius around o_b. The distance between objects is measured by Euclidian distance.

$$N_\varepsilon(o_b, C_i) = \{o_s | o_s \in C_i \wedge dist_{euclid}(o_b, o_s) < \varepsilon\} \quad (6)$$

Based upon $N_\varepsilon(o_b, C_i)$ we define the distance between an object o_b and a class C_i:

$$dist(o_b, C_i) = 1 - \frac{|N_\varepsilon(o_b, C_i)|}{|C_i|} \quad (7)$$

Thus the minimal distance is 0, if all objects of one class lie within the ε-neighborhood of o_b. The maximal distance is 1, if no object is inside ε-neighborhood. Equation 8 defines the class with the minimal distance to o_b among all classes.

$$cl_{rbc}(o_b, C) = \{C_i \in C | \forall C_j \in C : dist(o_b, C_i) \leq dist(o_b, C_j)\} \quad (8)$$

For $|cl_{rbc}(o_b, C)| = 1$ the RBC assigns o_b to the next class C_i. If $|cl_{rbc}(o_b, C)| > 1$, this means there is more than just one class with a minimal distance. Then one of these classes with a minimal distance is chosen at random.

6 EXPERIMENTS

In this section we evaluate the presented idea of video classification. So as to show the robustness of the approach against varying camera settings, the evaluation focuses on aspects of invariance. First rotation invariance with 9 different camera angles is analyzed. Then scale invariance with different zooming factors is tested. Third translation invariance is considered by shifting objects with repeating movements. The fourth subsection deals with invariance regarding time shift of an activity.

Test series are performed by own and by external video data. Own videos are recorded especially for the evaluation phase and external video data is taken from the online video database *youtube.com* (YouTube, LLC). In addition experiments with own video data are computed by m-fold cross validation. For classification process we use 10 classes, where each class consists of 20 videos (total 200 videos). External videos are analyzed by assigning them to especially recorded video classes, because cross validation was not possible due to classes with just few clips (total 102 videos). Further on in subsection 6.2 we classify 240 self recorded video sequences from different camera angles. Each video shows one of the next 10 home improvement activities: filing, hammering, planing, sawing, screwing, using a paint roller, a paste brush, a putty knife, sandpaper and a wrench.

6.1 Motion Transformation and Reflection Invariance

Next in fig. 5 an example 1D-function and its frequency domain is illustrated. Above one can see a 1D-function, which regards to the x-axis coordinate of centroids. This function captures the mean motion inside an external video clip. The motion in this clip arises from a person handling a wrench. Furthermore the 1D-function corresponds to the movement of the person, since the centroid moves from left to right and vice versa.

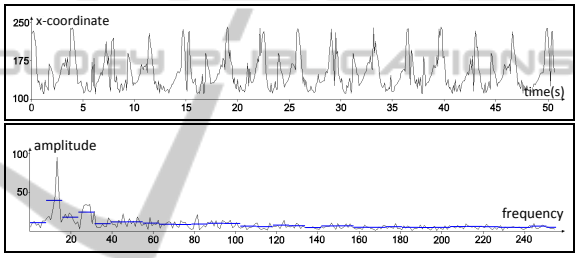

Figure 5: FFT of a 1D-function: Above 1D-function of a person handling a wrench, bottom FFT of this action.

The second plot below depicts the transform to the spectral domain. A partitioning of the frequency axis into $m = 32$ intervals leads to 32 AAFIs. Moreover the entirety of all AAFIs captures the mean information of the spectral domain without considering each single unit. Each significant frequency high or low has an influence on concerning AAFI. Moreover wide ranges with constantly high or low amplitudes are all captured by AAFIs and resulting feature vectors, too.

Considering fig. 5 it becomes obvious that this method provides horizontal, vertical and diagonal reflection invariance, because a reflection has no effect on the frequency of motion along one axis.

6.2 Spatial Rotation Invariance

Our next three test series focus on rotation invariance of the presented classification method. For each test series raw moments (centroids) are utilized. Videos from 9 different camera angles are classified. Except videos recorded from a frontal point of view 30 videos for each angle are assigned to one of ten classes. Each of these classes consists of 20 videos recorded from a

frontal view (total 440 videos in database). Frontal view videos are assigned by m-fold cross validation.

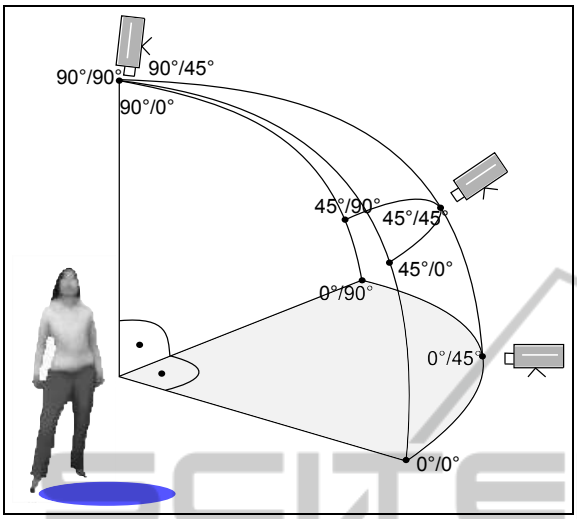

Figure 6: Illustrating camera angles.

The bar chart in fig. 7 depicts experimental results for especially recorded video data using directional information of image moments. In addition to it AAFI interval sizes are set to 8.

Videos recorded from a frontal point of view achieve a maximal accuracy at 0.87. The higher the horizontal and vertical camera angle, the lower the accuracies. The lowest accuracy is marked at 0.23 for a 90°/90° angle. This behavior is related to the fact, that the referenced classes contain only videos with a frontal camera position. In addition a frontal point of view gives clearer motion. Nevertheless 7 out of 9 camera angles achieve at least an accuracy of 0.40 and the average accuracy is 0.48.

This means our approach works even, if we rotate the point of view. There are two main reasons for this observation: First, if the angle is enlarged along just one axis, motion along the other axis stays almost unchanged. So for the motion feature vector of one axis there are little changes. Second, even if the camera angle changes, the frequency of a movement stays the same. Only the clearness of the motion direction descends.

Fig. 8 shows experimental results for own video data using position information of image moments. Here AAFI interval sizes are set to 4.

Frontally recorded videos result in a maximal accuracy of 0.80. Accuracies for each angle are varying strongly and the overall accuracy falls to 0.43. The lowest accuracy measured is 0.21 for a 90°/90° angle. Altogether 5 out of 9 camera angles give an accuracy of at least 0.40. In contrast to directional information of moments the position of moments is much

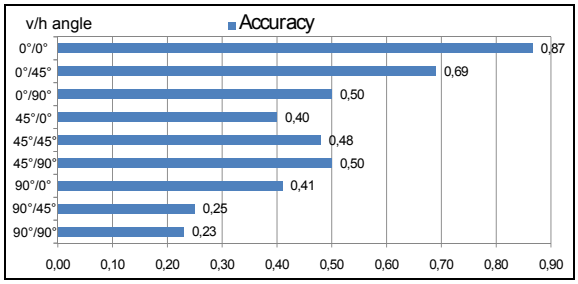

Figure 7: Accuracies of tests with raw moments and directional data.

more sensitive to camera angles, because the range of a movement affects directly the average amplitude of each frequency interval.

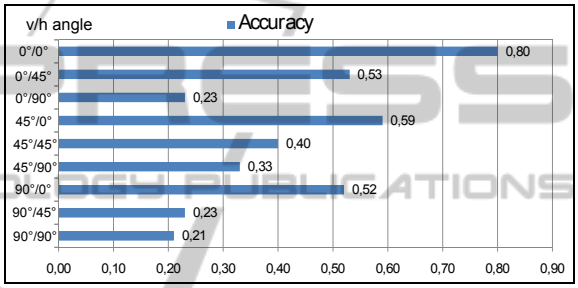

Figure 8: Accuracies of tests with raw moment positional data.

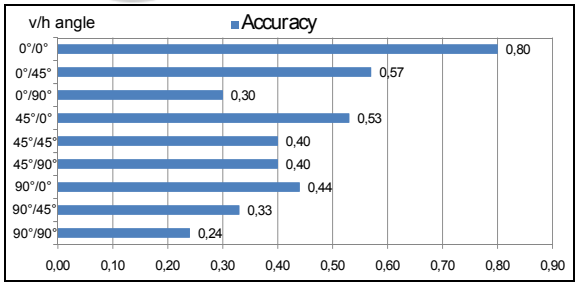

Figure 9: Accuracies of tests with raw moment positional data and normalized frequency domain.

Settings for tests regarding fig. 9 are the same like for fig. 8. The only difference is that we normalize here frequency values for classified clips as well as for referenced clips. Normalization is realized by dividing each frequency by the frequency maximum of the whole frequency spectrum. Thereby AAFIs of classified and referenced clips stay at the same level, even if the camera angle changes.

Here experimental results do not vary as strong as in fig. 8 and the average accuracy ascends to 0.45. 6 out of 9 camera angles yield an accuracy about 0.40.

6.3 Scale Invariance

Next two line charts present experimental results for scale invariance of our approach. The first line chart shows results for tests with own videos and the second line chart regards to external videos. Both internal and external video sequences are analyzed via raw moment positions (centroids), since moment directions are always scale invariant. This is related to the fact that a direction can only be -1, 0 or 1 (see 5). Scaling has no effect on this values. Experiments are conducted for 10 different scale factors beginning at 0.25 and ending at 4.0.

Fig. 10 illustrates how accuracies decline when scaling factor decreases or increases. By decreasing the zooming factor accuracies fall faster than by zooming in, because the clearness of a motion depends on the range, too. A zooming factor of 1.5 achieves an accuracy of 0.74 and a factor of 0.67 achieves 0.59. For zooming factors 0.5 and 2.5 accuracies stay above 0.30. So the approach works even for position information of raw moments as far as the zooming factor is not too high or too low. Normalizing frequency spectra by maxima or averages of each spectrum leads to constant accuracies 0.80 and 0.74. Raw moment directions give a maximal constant accuracy for all scaling factors at 0.87.

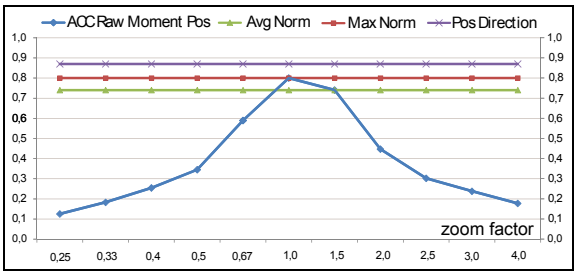

Figure 10: Accuracies for internal videos and different zooming factors.

Now in fig. 11 we see the same effect as in fig. 10. Accuracies decline when scaling factor decreases or increases. There is just one exception for scaling factor 1.5. For this factor accuracy increases from 0.40 to 0.42. This behavior is associated with the referenced classes. External videos are assigned to own video classes, where distances between camera and moving object in external clips are bigger than in own clips. Hence a zoom in aligns external and referenced AAFIs. A normalization of frequency spectra by maxima or averages of each spectrum results in constant accuracies 0.32 and 0.28. By utilizing moment directions the accuracy for each test series stays at 0.30.

Once again one can see that a zoom out has a stronger effect on accuracy descend than a zoom in, because motion ranges become smaller. Here this effect becomes even more apparent than in fig. 10, because external videos reveal more irregular motions.

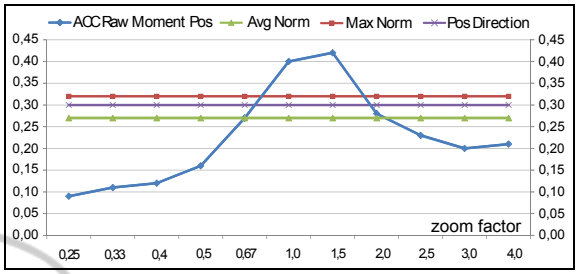

Figure 11: Accuracies for external videos and different zooming factors.

6.4 Translation Invariance

Varying positions of one activity in different videos do not influence classification process (translation invariance). But shifting motion areas within one video influences classification process. Next fig. 12 and 13 illustrate how accuracies change in this case. For each classified clip translation takes place frame by frame. Further on fig. 12 and 13 plot test results for different shift directions and shift velocities. Tests with own videos use directional information of moments and tests with external videos are performed by position information.

Fig. 12 visualizes how accuracies for own videos decrease, when translation velocity is increased. If vertical or horizontal shift of motion is realized, accuracies decrease slightly from 0.87 to 0.75 respectively 0.72. By contrast if diagonal translation is realized there is a strong decrease from 0.87 to 0.16. The explanation for this different behavior is that shifting a centroid along just one axis modifies just one coordinate. Unmodified coordinates result in unmodified features. The yellow line in fig. 12 depicts the accuracy when central moments (variance) are implemented. Here for each translation type and velocity the accuracy stays always at 0.81.

Next test series in fig. 13 shows that external video data reacts very sensitive on translation. At the beginning there is a abrupt descent for each curve. Then horizontal and vertical translation curves stay constantly at 0.27 and 0.23. Accuracy curve for diagonal shift ends at 0.17. For these abrupt descents two reasons can be stated: One reason is that external videos depend much more on just one 1D-function than own videos. Another reason is the greater sensitivity of positional information of moments to translation than directional information. Again central moments result in constant accuracies at 0.30 no matter which

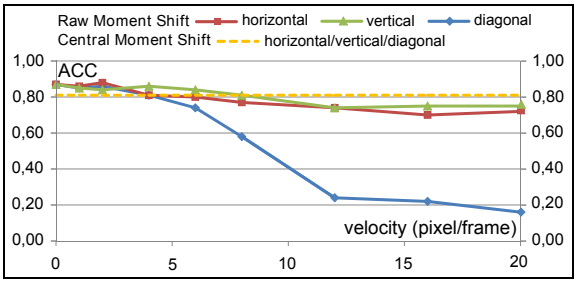

Figure 12: Test series with own videos and moments with translation.

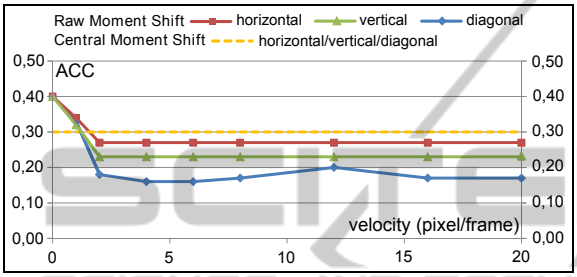

Figure 13: Test series with external videos and moments with translation.

translation type or velocity is applied.

Above experiments point out that video sequences with moving objects or moving cameras can often be classified more accurate with central moments than with raw moments. It should be taken into account that sequences recorded with moving camera need background subtraction.

6.5 Time Shift Invariance

Now we focus on time shift invariance of AAFIs. In this context time shift means, that the analyzed video starts at different points of time. In order to obtain regular shifts, we use sliding windows with a window size of 256 frames. The full length of a video is 512 frames. The window is shifted along the time axis stepwise. After each shift the action inside the sliding window is classified. Fig. 14 illustrates this idea for a 512 frames long video sequence.

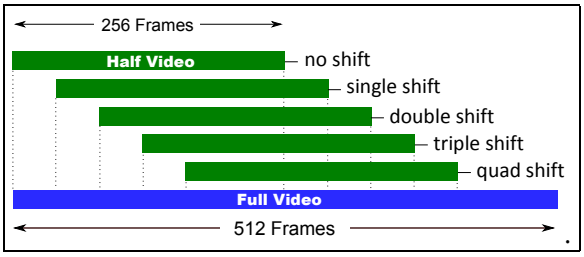

Figure 14: Time shift illustration.

This technique is used for internal and external videos during classification stage. Again internal videos are classified via raw moment directions and external videos are assigned by raw moment locations. In fig. 15 it becomes obvious, that the starting point of a repeating movement has only little effect on frequency spectra and resulting feature vectors. Here 256 frames of 512 frames long clips are shifted along the time axis and classified. Each shift has a length of 10 frames. Own videos stay for each shift around an accuracy level of 0.80 and external videos stay around 0.30. Hence the approach is almost invariant with regard to time shift.

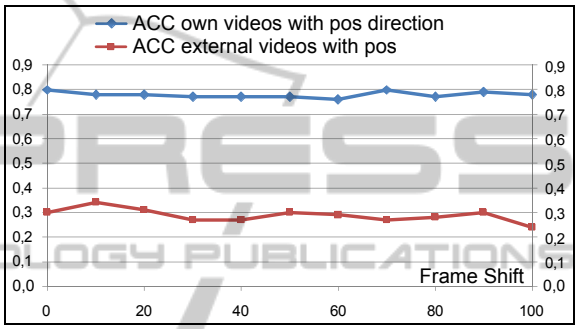

Figure 15: Accuracies for videos with different starting times.

7 RELATED RESEARCH

Video annotation and classification can be realized in many different ways. Main techniques base on key-frames (Pei and Chen, 2003), texts in frames (Lienhart, 1996), audio signals (Patel and Sethi, 1996) and motions. Each technique has to be robust against different disturbing factors. Focussing on motion and action recognition robustness against rotation and translation is an important task.

Translation invariant methods for human action classification can be found in (Fanti et al., 2005; Bobick et al., 2001; Niebles and Fei-fei, 2007), where approaches of Fanti et al. and Bobick et al. also fulfill scale invariance. A bulk of literature refers to rotation invariant motion classification. In (Chen et al., 2008) and (Bashir et al., 2006) rotation invariant methods for motion trajectory recognition are presented, where (Chen et al., 2008) provides only planar rotation invariance. Results in (Bashir et al., 2006) resemble our test results, but the maximal number of classes is set to 5 and the maximal angle size is 60°.

Further on some research work provides methods with rotation and scale invariance at the same time. Papers (Weinland et al., 2006; Rao et al., 2003) are based on Motion History Volumes respectively Mo-

tion Trajectories, whereas (Abdelkader et al., 2002) utilizes self-similarity plots resulting from periodic motion. Unfortunately the research work of Weinland et al. and Rao et al. do not analyze rotation invariance satisfactorily. The approach of Abdelkader et al. achieves high accuracies for a wide range of different camera angles. For a 1-nearest neighbor classifier and using normalized cross correlation of foreground images 7 out of 8 angles have an accuracy above 0.60. A comparison of this work to our work is not possible due to the fact, that Abdelkader et al. consider only one class for their classification process.

He and Debrunner compute Hu moments for regions with motion in each frame (He and Debrunner, 2000). Afterwards their system counts the number of frames until a Hu Moment repeats and define this number as the motion's frequency. Hu moments are invariant for translation, planar rotation, reflection and scaling. Here the periodic trajectory of an object cannot be ascertained. A strongly related work to our approach is given by (Meng et al., 2006). This paper depicts a time shift invariant technique for repeating movements, but it depends on the MLD-System (Moving Light Displays).
If we compare our approach to other approaches we find out, that other approaches do not comprise all different types of invariance as entirely as our method.

8 CONCLUSIONS

In this paper we have shown a scale, view, translation, reflection and time shift invariant approach for classifying video sequences. The classification process is performed by AAFIs, which represent average amplitudes of frequency intervals. Frequency spectra are figured by transforming spatio-temporal image moment trajectories via FFT. In addition a novel radius based classifier (RBC) was proposed, which improved the performance of the system. The stated accuracies in the experimental phase result from both selected features and RBC. Other classifiers (k-nearest neighbor, bayes, average link) we tested do not achieve same accuracy levels as RBC.

The system's robustness against different camera properties (zoom, angle, slide, pan, tilt) is useful for classifying clips from varying sources. Furthermore it stays an open issue to adapt and analyze the presented approach for real time action recognition.

REFERENCES

Abdelkader, C. B., Cutler, R., and Davis, L. (2002). Motion-based recognition of people in eigengait space. In *Proceedings of the Fifth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 267–277.

Ayyildiz, K. and Conrad, S. (2011). Video classification by main frequencies of repeating movements. In *12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011)*.

Bashir, F., Khokhar, A., and Schonfeld, D. (2006). View-invariant motion trajectory-based activity classification and recognition. *Multimedia Systems*, 12(1):45–54.

Bobick, A. F., Davis, J. W., Society, I. C., and Society, I. C. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267.

Chen, X., Schonfeld, D., and Khokhar, A. (2008). Robust null space representation and sampling for view-invariant motion trajectory analysis. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–6.

Fanti, C., Zelnik-manor, L., and Perona, P. (2005). Hybrid models for human motion recognition. In *IEEE International Conf. on Computer Vision*, pages 1166–1173.

He, Q. and Debrunner, C. (2000). Individual recognition from periodic activity using hidden markov models. In *Workshop on Human Motion*, pages 47–52.

Lienhart, R. (1996). Indexing and retrieval of digital video sequences based on automatic text recognition. In *Fourth ACM int. conf. on multimedia*, pages 419–420.

Meng, Q., Li, B., and Holstein, H. (2006). Recognition of human periodic movements from unstructured information using a motion-based frequency domain approach. *IVC*, pages 795–809.

Niebles, J. C. and Fei-fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Patel, N. and Sethi, I. (1996). Audio characterization for video indexing. In *SPIE on Storage and Retrieval for Still Image and Video Databases*, pages 373–384.

Pei, S. and Chen, F. (2003). Semantic scenes detection and classification in sports videos. In *Conf. on Computer Vision, Graphics and Image Proc.*, pages 210–217.

Rao, C., Gritai, A., Shah, M., and Syeda-Mahmood, T. (2003). View-invariant alignment and matching of video sequences. *IEEE International Conference on Computer Vision*, pages 939–945.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. In *Computer vision and image understanding*, pages 249–257.

Wong, W., Siu, W., and Lam, K. (1995). Generation of moment invariants and their uses for character recognition. *Pattern Recognition Letters*, 16:115–123.

Zhang, R., Vogler, C., and Metaxas, D. (2004). Human gait recognition. In *Proc. of the 2004 Conf. on Computer Vision and Pattern Recognition*, pages 18–27.