VIDEO CLASSIFICATION BY MAIN FREQUENCIES OF REPEATING MOVEMENTS

Kahraman Ayyildiz, Stefan Conrad

Heinrich Heine University Duesseldorf
Department of Databases and Information Systems
Universitätsstraße 1, 40225 Duesseldorf, Germany

ABSTRACT

This paper discusses an approach, which allows classifying videos by frequencies. Many videos contain repetitive activities like walking, swimming, or playing table tennis. These activities usually repeat with a certain frequency, which can be seen as a feature of cyclic motion. So determining this feature can help to classify videos with repeating movements properly. In this paper we explain how to find out the right frequencies for video clips and how to use them for classifying. The main method handles series of image moments as a function in order to transform this function into the frequency domain via FFT. Techniques proposed in this paper are tested with own and with external video data. Thus it can be pointed out how the system handles different data types and data qualities.

1. INTRODUCTION

Classifying videos is an important task in many branches. Multimedia or video databases can be found in archives of major corporations, governmental institutions, media branches, or museums. Furthermore the World Wide Web includes millions of video clips. Online video stores or online video portals, where people can upload their own clips, are two examples for online video databases. All of these sectors need efficient algorithms to classify the amount of existing clips or videos automatically.

Provider sided classification is expensive. User sided classification has low costs, but the quality of indexing is not ensured. Both approaches commonly use general descriptions for videos, which capture just the mean content. Nevertheless there are activities in clips which are not captured by these descriptions. An automatic indexing and annotating process could remedy these deficits. There are many techniques for classifying videos automatically, but one technique is barely investigated: Assigning videos by frequencies of repetitive movements.

In this paper we extract frequency features from periodic motion in videos. For this purpose regions of movement are captured frame by frame. At the same time image moments for these regions are calculated. A series of image moments represents a function, which assigns one moment to each frame of a video. This function again is transformed via Fast Fourier Transformation (FFT) and spans a frequency spectrum. The frequency spectrum reveals high amplitudes at certain frequencies. Combining these frequencies gives a multidimensional feature vector for each clip. By integrating a classifier these feature vectors can be used during the classification process.

In the next section some related work to our approach is introduced. Then the process of extracting features from videos and utilizing them for classifiers is discussed in section 3. Afterwards we describe how to derive so-called *1D-functions* from image moments in section 4. Later on a radius based classifier, which is used for our approach, is introduced and explained in section 5. In section 6 the idea of classifying videos by frequencies is evaluated with home improvement videos. In the last section our approach for assigning video clips is reviewed.

2. RELATED WORK

Video annotation, classification and retrieval can be realized in many different ways, because videos reveal a huge amount of information. Single frames [1], text in frames [2], audio signals [3], or motion in videos can be used for feature extraction.

The most research work related to motion recognition focuses on the gait or the gestures of humans. An approach for recognizing human motions in general is delivered by [4]. The authors utilize the flow and the strength of change of pixels as features. Some research concentrates on cyclic motion in clips. In [5] the periodical movement of human body parts is captured by Moving Light Displays (MLD). Pieces of curves described by these MLDs are used as reference curves. Another method related to cyclic motion recognition is proposed by He and Debrunner [6]. Calculating Hu Moments for regions with motion in each frame, they count the number of frames until a Hu Moment repeats and define this number as frequency. A widely cited work on periodic movement recognition originates from Polana and Nelson [7]. They divide each frame of a clip into 16 parts of same size and store 6 frames for each repeating motion. Then pixel activities for these 6 x 16 parts are measured and summed up for every cyclic motion. The resulting 96-dimensional feature vector is used for classifying.

Strongly related works to our frequency domain based approach are [8], [9], and [10].

3. CLASSIFYING VIDEOS BY FREQUENCIES

In this section we discuss methods for classifying videos by their frequencies. That means videos with repeating movement sequences like hammering, planing, or filing are considered. It is even possible to extend our approach to other topic areas like sports (tennis) or music (accordion). The main requirement is always a clear and rather constant motion. Our approach is not able to classify activities with a strong variation of frequency spectra like playing a violin. Further periodic texture motion cannot be captured.

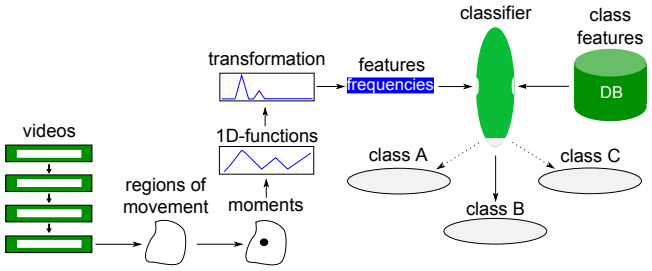

Fig. 1. Flow diagram of whole classification process

In figure 1 we present the different stages of our approach. As a first step of the whole classification process regions of movement are detected in every clip frame by frame. Regions are detected by measuring the color difference of pixels in two frames (current and previous frame). With these regions image moments can be figured. In the following approach two types of moments are applied: centroids and pixel variances (see section 4.2). From these moments 1D-functions can be derived, which represent the motion in a clip. These functions again are transformed into the frequency domain via FFT. The frequencies with the highest amplitudes inside the frequency spectrum are considered as feature vectors for a clip with cyclic motion. After resolving the feature vectors a classifier can decide to which class a video fits best by comparing the features with features of other clips stored in database.

4. IMAGE MOMENTS AND 1D-FUNCTIONS

In order to compute the frequency spectrum for a video scene the motion has to be localized for each frame. In addition image moments are needed for calculating 1D-functions. So this section explains how to detect regions of motion and how to derive 1D-functions from these regions.

4.1. Regions of Movement

Figure 2 illustrates how we detect regions with motion by analyzing one of our clips. It shows two consecutive frames with a person troweling a wall. The color differences between these frames are measured for each pixel. If the color difference of a pixel is above a predefined threshold and if there are enough neighbor pixels with a color difference beyond the same threshold, then this pixel is considered as a pixel which is part of a movement. Hence a region of movement is the affiliation of pixels with movement.

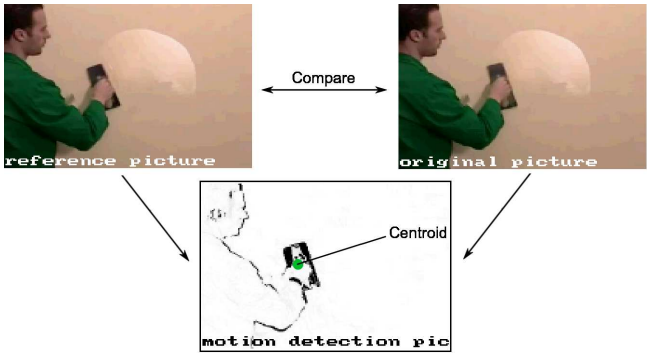

Fig. 2. Regions with pixel activity and centroid

The binary image below the two frames compared shows the regions with motion. A further interesting aspect of this illustration is, that the centroid of regions with motion lies exactly on the right hand. This means the centroid follows the movement of the troweling.

4.2. Image Moments

An image moment is a weighted average of pixel intensities of a picture. It can describe the area, the bias or the centroid of segmented parts inside a picture. There are two main types of image moments: raw moments and central moments. The difference between these two moment types is, that central moments are translational invariant and raw moments are not. For a two dimensional (grayscale) image $b(x, y)$ and $i, j \in \mathbb{N}$ a raw moment M_{ij} is defined as follows [11]:

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

M_{ij} is always of the order $(i + j)$. For a given binary image function $b(x, y)$ the area of segmented parts is determined by M_{00}. By $(\bar{x}, \bar{y}) = (M_{10}/M_{00}, M_{01}/M_{00})$ the centroid of segmented parts is defined. Applying centroid coordinates central moments can be computed by equation 2 [11].

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

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

4.3. Deriving 1D-functions

We define a 1D-function as a series of one-dimensional moment values. The series is arranged in a chronological order and corresponds to the sequence of frames in a film. This definition leads to the function $f(t)$ with t as time. Hence $f(t)$ represents cyclic motion along one axis. Let $(\bar{x}_t, \bar{y}_t) = (M_{10_t}/M_{00_t}, M_{01_t}/M_{00_t})$ for the coordinates of a centroid regarding to time t. Then function $f_c(t) = (\bar{x}_t, \bar{y}_t)$ implies:

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

In section 6 $f_{c_x}(t)$ and $f_{c_y}(t)$ are used for experimental test series instead of $f_c(t)$, because transforming 1D-functions results in better accuracies than transforming 2D-functions. For the same reason two separate 1D-functions of central moments are implemented and tested:

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

For any 1D-function $f(t)$ we define the speed of an image moment at time t as follows:

$$f_s(t) = |f(t) - f(t-1)| \tag{5}$$

The direction of a moment at time t is defined by equation 6.

$$f_d(t) = \begin{cases} +1, & \texttt{if } f(t) - f(t-1) > 0 \\ 0, & \texttt{if } f(t) - f(t-1) = 0 \\ -1, & \texttt{if } f(t) - f(t-1) < 0 \end{cases} \tag{6}$$

5. RADIUS BASED CLASSIFIER

Now we introduce a classifier, which turned out as very effective during our experimental stage. We name our classifier *Radius Based Classifier* (RBC), because the radius around a tested object has an important function. For a given object o and a radius ε the RBC counts all objects of a class C_i around o with a distance smaller than ε. The normalized sum of all objects leads to a distance $dist(o, C_i) = 1 - \frac{|N_\varepsilon(o,C_i)|}{|C_i|}$ between tested object and class. After computing distances to all existing classes, the RBC assigns o to the class with the smallest distance.

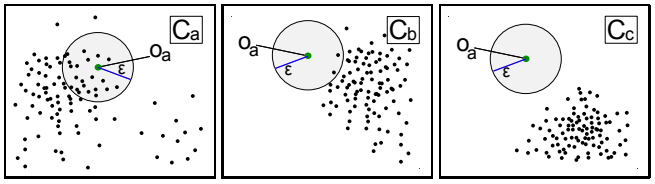

Fig. 3. Classifying with RBC

Figure 3 illustrates how the RBC works: An object o_a of an unknown class has to be classified. Therefore it is assigned to each existing class in order to calculate the class with the smallest distance. There are three different classes C_a, C_b and C_c, where each class has its own typical object distribution. Assigning o_a to class C_a reveals, that there are many objects within radius ε. In class C_b only 2 objects are present inside the given metric. Objects of class C_c are far away from o_a, so there is no object of this class within radius ε.
According to these three classes, o_a fits best into class C_a, because it is part the typical object distribution. At the same time this fact leads to a minimal distance.

6. EXPERIMENTS

This section discusses test series based upon the presented idea of video classification. First, experiments regarding to moment type, moment speed, and moment direction are considered. Second, translational invariance of motion classification is discussed and analyzed. In both subsections firstly test series with own video data and secondly test series with external video data from the online video database *youtube.com* are realized [12]. Furthermore tests with own video data are calculated by m-fold cross validation. 10 classes, where each class consists of 20 videos, are tested (total 200 videos). External video data is tested by assigning clips to own classes, because cross validation was not possible due to classes with just few clips (total 102 videos). All videos show following 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. Raw Moments and Central Moments

In figure 4 an example of a 1D-function and its transformation is illustrated. The upper plot shows a 1D-function of a clip with a person using a wrench. This function regards to the x-axis coordinate of centroids. One can see, that the centroid moves from left to right and vice versa, which corresponds to the movement of the person. Below this 1D-function its transformation to the spectral domain is plotted, where two maxima can be figured out. Both frequencies at these two maxima are used as feature vectors for this video clip during the classification process. In our approach we use up to three maxima for motion along each axis, if each maximum exceeds the average frequency clearly.

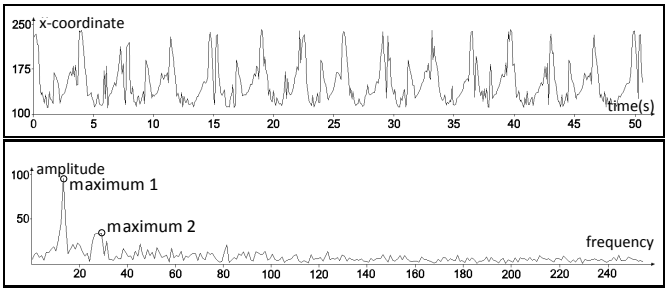

Fig. 4. FFT of a 1D-function: Above 1D-function of a person handling a wrench, bottom FFT of this action

The following two bar diagrams in figure 5 focus on results of test series with raw and central moments. Moreover results of tests with directional and speed information of moments are included. Experiments need a tuning of parameter ε for the RBC since different moment types and data sources result in varying features. The left diagram refers to tests with video data, which was produced especially for our experiments. The right diagram relates to external video data from an internet database [12]. At first glance it becomes apparent, that the raw moments respectively centroids result in better accuracies in almost all cases. There is only one exception for speed information of central moments respectively variances with external data. Further own videos can be classified much mo-

re efficient than external videos. This is associated with the fact, that the external videos have lower quality in the sense of regular movement, camera positioning and scaling. For recorded video data and directional information of centroids our approach achieves a maximal accuracy of 0.70. Experiments with external videos and centroid coordinates result in a maximal accuracy of 0.40. Here centroid coordinates achieve higher accuracies than centroid directions, because external videos contain more irregular movements. For both data sources the speed information of centroids is a weak feature for classifying.

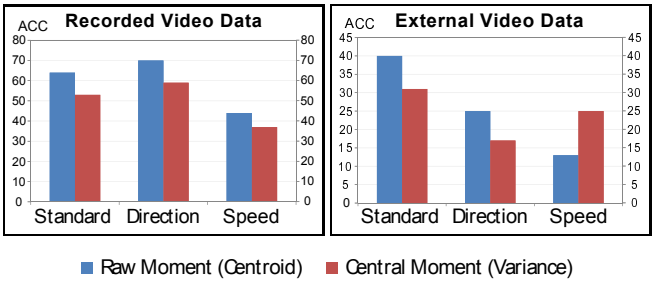

Fig. 5. Accuracies of tests with raw and central moments

The accuracies result from both selected features and RBC. In further test series, which are not listed in this work, we compared the RBC with other classifiers. We detected that the RBC improves results, but does not affect the relative highs of feature accuracies.

6.2. Translational Invariance

Different positions of one activity in different videos have no effect on classification process (translation invariance). Figure 6 shows how accuracies change, when motion areas are shifted within one video. The translation takes places for each classified clip frame by frame. Furthermore tests with different shift velocities and shift directions are plotted. Again own and external video sources are integrated. Tests with own videos are performed via directional and tests with external videos are realized via standard information. For own videos and centroids the accuracy decreases constantly with increasing velocity of translation. Moreover accuracies of a diagonal translation decrease much faster than accuracies of horizontal or vertical translation, because shifting a centroid along just one axis does modify just one coordinate. Unmodified coordinates result in unmodified feature vectors. The yellow line shows the accuracy for central moments (variance). For each translation type and velocity the accuracy stays constantly at 0.59. Considering test series with external data, it becomes apparent, that accuracies react very sensitive on translation. At the beginning each curve falls rapidly and then decreases constantly. There are two reasons for this behavior: First external videos depend much more on just one 1D-function and second tests with standard moments are more sensitive to translation than directional information of moments. On the other side here central moments lead to constant accuracies, too. For any translation type and velocity the accuracy is 0.31. According to these experiments it can be stated, that clips with moving objects or moving cameras can often be classified more accurate with central moments than with raw moments.

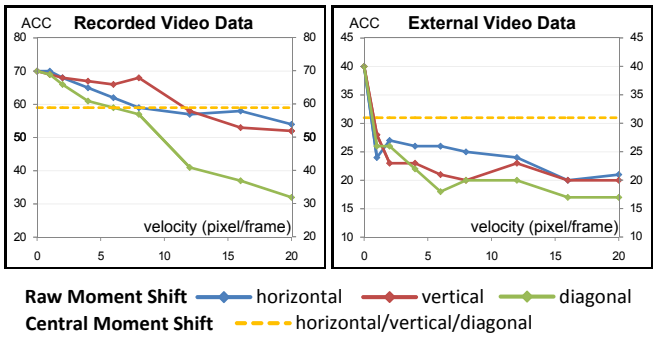

Fig. 6. Accuracies for moments with translation

7. CONCLUSION

In this paper we have shown a novel approach for classifying clips via main frequencies of repetitive movements. These frequencies can be figured out by transformed 1D-functions of image moments. Beside different 1D-functions we defined, we explained a novel radius based classifier for our purpose. The experimental stage exposed, that our approach works accurately for centroids as image moments. But for videos including translation of motion translationally invariant central moments work more efficient. A further aspect, which remains for future research, is the classification of clips by intervals of frequency spectra.

8. REFERENCES

[1] S.C. Pei and F. Chen, "Semantic scenes detection and classification in sports videos," in *Conference on Computer Vision, Graphics and Image Processing*, 2003, pp. 210–217.

[2] R. Lienhart, "Indexing and retrieval of digital video sequences based on automatic text recognition," in *Fourth ACM international Conference on Multimedia*, 1996, pp. 419–420.

[3] N. Patel and I. Sethi, "Audio characterization for video indexing," in *SPIE on Storage and Retrieval for Still Image and Video Databases*, 1996, pp. 373–384.

[4] R.V. Babu and K.R. Ramakrishnan, "Compressed domain human motion recognition using motion history information," in *ICIP03*, 2003, pp. 321–324.

[5] P. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection," in *Pattern Recognition*, 1994, pp. 1591–1603.

[6] Q. He and C. Debrunner, "Individual recognition from periodic activity using hidden markov models," in *Workshop on Human Motion*, 2000, pp. 47–52.

[7] R. Polana and A. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal of Computer Vision*, vol. 23, pp. 261–282, 1997.

[8] Q.G. Meng, B.H. Li, and H. Holstein, "Recognition of human periodic movements from unstructured information using a motion-based frequency domain approach," *IVC*, pp. 795–809, 2006.

[9] Fangxiang Cheng, William Christmas, and Josef Kittler, "Periodic human motion description for sports video databases," *International Conference on Pattern Recognition*, vol. 3, pp. 870–873, 2004.

[10] Ross Cutler and Larry S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[11] W. Wong, W. Siu, and K. Lam, "Generation of moment invariants and their uses for character recognition," *Pattern Recognition Letters*, vol. 16, pp. 115–123, 1995.

[12] YouTube, "Youtube: Broadcast yourself," www.youtube.com.