An Object-Behaviour Based Key-Frame Extraction Method

Ziran Wu*, Bumeng Liang, Guichu Wu

Key Laboratory of Low-voltage Apparatus Intellectual Technology of Zhejiang, Wenzhou University, Wenzhou, 325035, China

***Abstract* — This paper introduces a novel object-behavior based key-frame extraction method which employs the HOG method to locations of target objects, and the GDT&OM method to match the frames that can illustrate video clips best. In the experiments, human action video clips are used as examples. A conventional key-frame extraction method called PME is implemented as a comparison. The final results indicate that the proposed method can extract more 'meaningful' key-frames than the conventional method.**

Keywords - key-frame extraction; human behavior; human detection; template matching

I. INTRODUCTION

Motion analysis methods for sports in prior art usually segment objects by their motion information. Matter et al. [1] developed a hybrid system that analyze footages footage to detect cuts and classify camera movement. High-level semantic information was manually selected as logs. Liu et al. [2] proposed a method extracting key-frames by analyzing motion factors, such as moving velocities or accelerations of image regions, between frames. Then the key-frames were selected as logs of a video. The paper [3] presented Luo et al. illustrated a method that detected differences in a video between two continuous frames to extract moving objects, abstracted the video into a number of clusters, and used a Bayesian Network to analyze the motion pattern. Li et al. presented a method [4][5] that for each clip extracts human bodies by estimating motion energy between two frames, selected frames that have motion energy larger than a threshold as key-frames, and finally used an HMM to recognize athletes' actions. Our method, however, aims to select a frame which containsan object being able to indicate the video shot best.

[1] In the experiment, human bodiesare used as target objects. The method selects a human pose represent the human action best in every shot. Meanwhile, the action type of a frame sequence can also be determined because every selected key-frame reflects a class of action.

Of course, "key-frame" is a concept that has a lot of subjective factors. In other words, determining whether a key-frame is a success or failure strongly depends on individual opinions. In this paper, we also introduce a conventional key-frame extraction method as a comparison. We compare the results between this method and our method according to the rule that for a frame sequence the key-frame which can represent the content of a frame better is the winner.

II. KEY-FRAME EXTRACTION METHOD

Our system which is designed to process videos representing human actions, contains four procedures for extracting the key-frame from an input frame sequence:

Each frame is processed by Histogram of Oriented Gradients (HOG) human detector [6][7]to detect human bodies.

Detected human bodies (in fact these are image regions containing human bodies) are scaled to the same size as the prepared templates. Each human body is matched with the template set by Generalized Distance Transform and Orientation Maps (GDT&OM)[8]and the best match is found(smallest matching score). All frames' best matching scores are compared and the smallest is chosen. The corresponding frame is selected as the key-frame.

In the experiments, we use the Weizmann dataset [9] as the test dataset. This dataset contains videos of full human bodies without overlapping, which is suitable for our detection and matching methods. Each video in the dataset represents a single shot and a single person performing a single action type. Besides, it contains not only some typical common action types, such as walking and running, but some abnormal action types such as jacking and skipping as well.

[1] *Address correspondence to this author at Key Laboratory of Low-voltage Apparatus Intellectual Technology of Zhejiang, Wenzhou University, Wenzhou, 325035, P.R.China; Tel: +86 18857745619; Fax: +86 577 88373126; E-mail: nature.nano@gmail.com

DOI 10.5013/IJSSST.a.17.27.08

ISSN: 1473-804x online, 1473-8031 print

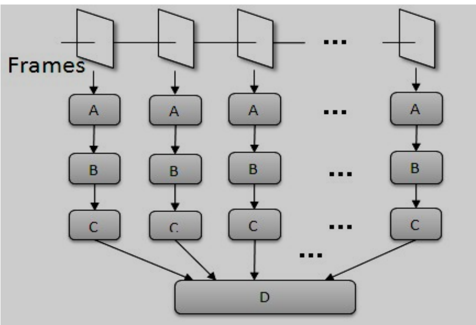

Fig.1 Diagram of the key-frame extraction system

A: HOG human detection; B: Scaling the detected windows to the same size of the templates; C: GDT&OM template matching to find the best match; D: select the frame with the smallest matching score as the key-frame

A. HOGhuman detection

We used an HOG human detector [7]to locate human bodies in frames. Firstly, an N×M image sample is converted to a gradient copy:

$$G(i,j)=\sqrt{G_h(i,j)^2+G_v(i,j)^2} \tag{1}$$

where (i, j)is the coordinate pair of a pixel, Gh and Gvindicate the vertical and horizontal gradient operators respectively. The orientation of each pixel is computed by:

$$O(i,j)=G_v(i,j)/G_h(i,j) \tag{2}$$

Then the gradient copy is divided into blocks, each of which contains 2×2 cells. Each cell is an 8×8 pixel region. In each cell a histogram is built. The bins of the histogram accord to the orientations and the height of each bin is the sum of gradients, which can be represented by:

$$H(k)=\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}G(i,j)F(O(i,j),k) \tag{3}$$

$$F(x,k)=\begin{cases}1, \text{ if } x\in R(k)\\ 0, \mathit{else}\end{cases}$$

Where H(k) is the height of bin k, and R(k) is the orientation range of bin k. The histograms are then dumped into a feature vector. A number of positive image samples (containing objects) and negative image samples (non-object) are processed to generate feature vectors and a two-class support vector machine (SVM) [10] is applied to train a classifier (positive samples for one class and negative samples for the other). An object detector is designed to apply regions of images to the classifier and determine whether a target objects is contained in each region.

The HOG detector is able to perform a high true detection rate against a low false rate, with a good compatibility for objects' different appearances.

B. Generalized Distance Transform and Orientation Maps

For each video clip, we select its key-frame by measuring the behavior of the target object. The behavior is reflated by a pose in each frame. So we try to find a key pose that can best indicate the behavior represented the entire video clip. Hence the frame contains the key pose is to be chosen as the key-frame.

We have made a number of contour templates each of which can significantly represent a type of action. Generalized Distance Transform is a template matching algorithm that computes distances of each pixel to the nearest edge. We generate a binary edge image by a Canny operator and perform GDT according to the Canny edges. The formula of GDT can be represented as follows according to Felzenszwalb and Huttenlocher[11]:

$$D_{\Psi}(p)=\min_{q\in\varsigma}\{d(p,q)+\Psi(q)\} \tag{4}$$

whered(p,q) is the Euclidean distance between point p and q:

$$d\left(p,q\right)=\sqrt{(p_x-q_x)^2+(p_y-q_y)^2}$$

Meanwhile ζis defined as a regular grid and Ψ(q) represents a function on the grid. Here Ψ(q) is defined as:

$$\Psi(q)=\begin{cases}\dfrac{1}{\sqrt{I_x^2+I_y^2}}, \text{ if } (q)\in e\\ \infty, \qquad \text{otherwise}\end{cases} \tag{5}$$

wheree is the binary edge pixels. Ix=∂I/∂xand Iy=∂I/∂y, are gradients in horizontal and vertical directions at point q in image I. Here Ψ(q) is different from the conventional distance transform (DT), which assigns Ψ(q)=0 when q□e. Equation 4 means that for a pixel p, find the nearest pixel q on edges, and assign the distance value as the Euclidean distance between p and q plus a small value based on its gradient.

According to Thanh's paper [8], the orientation value of a pixel is related to the position of its nearest edge pixel. That is to say, the orientation of edge pixels will impact the non-edge pixels nearby. Considering a single pixel p in image I, assuming that q* is the nearest edge pixel top (which means that orientation maps are discontinuous), the orientation value at pixel p is defined as

$$O_{\Psi}(p)=\arctan(I_{x*}/I_{y*}) \tag{6}$$

whereIx* andIy* are the gradients at q*.

III. EXPERIMENTS

A. HOG Detector

Training data consists of training images of the INRIA dataset [7]and part of human body windows from the Weizmann dataset[9], because the Weizmann dataset includes some abnormal poses. Finally we have 2812 positive training images. We also add negative samples which present parts of human bodies (e.g. arms, legs, torsos) to the initial INRIA negative sample set. So finally we have 19980 negative samples. This training scheme is able to significantly reduce false detections locating on parts of human bodies, which is an improvement we made upon the initial HOG detector scheme. The test data includes 9 persons and 10 action classes, totally 5016frames. Because the test images are simple (simple backgrounds, simple poses, no overlapping), we set the merging threshold t to 0.00, in order to achieve a maximum true positive detection number. Table I. shows the result.

TABLE I. HOG RESULT FOR THE WEIZMANN DATASET

Actions	Number of frames	Detected bodies	Detection rate
Jacking	729	691	94.79%
Jumping	458	419	91.48%
Running	405	397	98.02%
Skipping	494	413	83.60%
Walking	711	694	97.61%
Galloping	481	468	97.30%
Place-jumping	454	431	93.50%
One-waving	438	408	93.15%
Two-waving	445	423	95.06%
bending	401	337	84.04%
Total	5016	4671	93.12%

According to Table 1, most action classes can be well detected. The detection for skipping and bending images is the worst detection because skipping is an abnormal pose which has only a small number of training images. There are two action classes that the detector does not performed on well (skipping, bending). The reason is that there are not enough positive training images for these two classes, which makes them to be abnormal samples to the trained model.

B. GDT&OM Template Matching

Since there is no clear definition of what a key-frame of a sequence should be, selecting a frame that understandable for a human can also be considered as a choice. That is to say, people can understand what happens represented by a sequence only by seeing its key-frame. For applications such as movie producing and storyboard making, template matching can be a good approach. Our method is mainly applied for the sequences that contain moving human bodies.

We use the Weizmann dataset as the test dataset, which includes ten action classes. We select six of them which are running, walking, jumping, skipping, jacking and side moving, because these six classes are stronger action types. We select a set of templates from all poses of the six action classes. Each action class has a pair of templates to represent two different moving directions, except jacking, whose templates are all symmetrical. The selected templates are considered to be significantly different from each other so that each of them can represent its action class uniquely. In the experiment, we built the template set as Fig2.

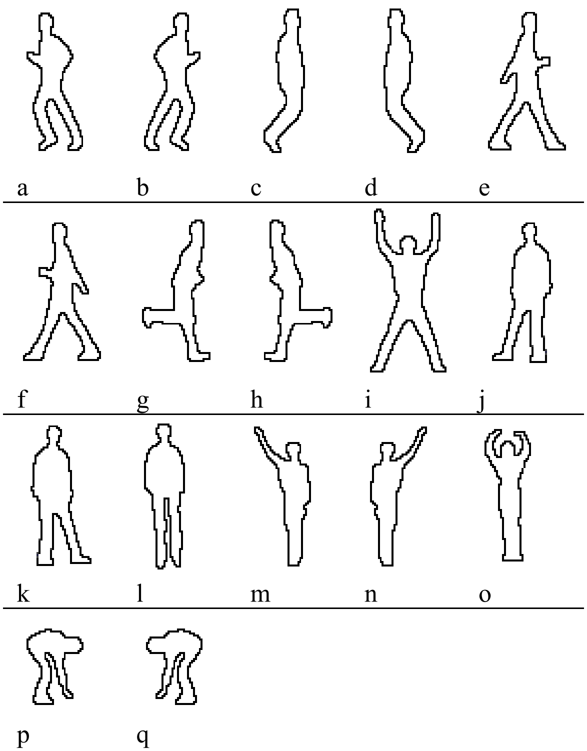

Fig.2 Template set for key frame selection

(a) and (b) running; (c) and (d) jumping; (e) and (f) walking; (g) and (h) skipping; (i) jacking; (j) and (k) galloping sideways; (l) jumping at a same place; (m) and (n) one-hand waving; (o) two-hand waving; (p) and (q) bending.

The detected image regions are scaled to the same size of the templates. In practice the detection windows have some margins between human bodies and window boundaries, but the templates don't, so the detection window will be slightly larger than the templates. We execute a quick scan inside the detection window by the templates. The reason why we do such procedure is that sometimes the body of a detection result is not just at the center position of the window due to different body poses and the merging process. If we used the same sized templates, the edges of the templates and the bodies in the windows would have a small dislocation, which leads to inaccurate matching results. If we employ smaller templates, scan inside the detection window, and finally select the best match, the result will be much more accurate. False detections have little influence on matching results, because these detections are far different from human body shapes and will get poor matching scores. Since the set of templates are only a part of all poses, it is pointless to discuss matching accuracy. What we want to see is that how well selected key frames represent their corresponding videos. The final result will be illustrated in Chapter 4.

IV. COMPARISON WITH PME

A. *Perceived Motion Energy*

Before we present the final result, we introduce a comparison method. We use a method based on perceived motion energy model (PME) proposed by Liu, Zhang and Qi [2] as the comparison method. This method implements a conventional idea that the system selects frames in which the objects have largest moving speed and smallest acceleration, in other words, the largest motion energy, as the key-frame. They calculated the PME of a frame by accumulating the motion vectors between two serial frames, and used a triangle model to determine the maximum PMEs.

According to the paper [2], PME is defined as “a combined metric of motion intensity and motion characteristics with emphasis on dominant motion”. In other words, in a video all motion is projected onto the dominant motion and the sum is computed as the measurement of motion energy. Since the PME method is based on MPEG video compression, each backward frame (B frame) in a MPEG video is segmented into a number of macro blocks with a certain size (e.g. 8×8 or 16×16). Each macro block is used to compute motion vectors and then to form motion vector fields. The magnitude of a macro block reflects the motion intensity of the block and the accumulation of the magnitudes can reflects the global motion intensity of the entire frame.

In Liu’s work, they represented PMEs in diagrams and perform a triangle mechanism to estimate peak PMEs. Then corresponding frames of these peak PMEs are selected as key frames. However, we aim to select a key frame for each shot, so we just select the frame with the largest PME of a frame sequence.

B. *Conclusion*

PCB is an important component for electronic devices - Mechanical connections and electrical transmission, thermal failure is its main failure mode, the heat flow characteristics research PCB thermal design is the basis and premise. Based on the principles of fluid mechanics, using the finite volume method for the thermal characteristics of the PCB were modeled to obtain the maximum junction temperature of the PCB, PCB's heat distribution and ambient temperature at different ambient temperatures on the PCB thermal characteristics the relationship for PCB thermal design provides a theoretical basis.

C. *Results and Comparison*

When processing a frame sequence, after we achieve the best matches for each frame, we select the frame with smallest matching score, in other words, the most similar to its matching template, as the key-frame. Meanwhile, the action class of the frame sequence can be determined as what the best matching template represents. Here is a comparison result between PME and our method in Figure 3. Only two class actions – walking and running, which have significant differences between the results for the PME method and our method – are presented. For other motion categories, the differences of the results between the two methods are very small, so the results are not listed below.

As argued above, deciding whether a key-frame is good or bad is subjective. According to Figure 3, most key-frames of the two methods are satisfying. However, under the rule that “a key-frame should represent the content of the frame sequence”, we argue that our method performs better for a few of these frame sequences.

For walking sequences, the key-frames extracted by the PME method appear to have different poses. While those extracted by our method can be considered as in the same pose. Some key-frames from template matching are easier to be recognized as walking, such as the sequences named ‘Ido walk’ ‘Lena walk1’ and ‘Lena walk2’.

For running sequences, the results between the two methods are closer. But there are still a few key-frames extracted by our method can be considered as better ones, such as ‘Moshe run’ and ‘Shahar run’.

TABLE II. “GOOD” KEY-FRAMESACHIEVED BY PME AND OUR METHOD.

Actions	PME	Ours	Total
Jacking	6	9	9
Jumping	8	7	9
Running	8	10	10
Skipping	8	8	10
Walking	7	10	10
Galloping	7	9	9
Place-jumping	9	9	9
One-waving	9	9	9
Two-waving	9	9	9
bending	7	9	9
Total	78	87	92

Table II. lists how many “good” key-frames achieved by both PME and our method. PME is more like a random selection in the view of frame content, and only performs better than our method for jumping images. Generally speaking, according to the table, our method achieves more understandable key-frames than PME. That is to say, the designing purpose of the key-frame extraction method is achieved.

V.CONCLUSIONS

Compared with a conventional method called PME, the entire key-frame extracting system achieves better experimental result. We argue that extracting key-frames representative for content of videos is more meaningful and useful in practice. We have tested the system by using human motion videos. Meanwhile, it can be simply extended to other types of objects by varying image samples. However, the method requires pre-definition of objects. A thought of further improvement is to develop a method that intelligently determine the primary content (e.g. the main object or event) of a video and determine the key-frame.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

REFERENCES

[1] J. Mateer and J. Robinson, “A vision-based postproduction tool for footage logging, analysis, and annotation,” Graphical Models, Vol. 67, Issue 6, p. 565–583, 2005.

[2] T. Liu, H.-J. Zhang and F. Qi, “A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model,” IEEE Transactions on Circuits and Systems for Video Technology, Volume 13, Issue 10, pp. 1006-1013, Oct. 2003.

[3] Y. Luo, T.-D. Wu and Jenq-Neng, “Object-based analysis and interpretation of human motion in sportsvideo sequences by dynamic bayesian networks,”Computer Vision and Image Understanding, p. 196–216, Volume 92, Issues 2–3, November–December 2003.

[4] H. Li, S. Lin, Y. Zhang and K. Tao, “Automatic Video-based Analysis of Athlete Action,” Image Analysis and Processing, pp. 205-210 , 2007.

[5] H. Li, S. Lin, Y. Zhang and K. Tao, “Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences,” Circuits and Systems for Video Technology, IEEE Transactions, pp. 351-364, March 2010.

[6] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Conference on Computer Vision and Pattern Recognition, pp. 886 - 893, 2005

[7] N. Dalal, "Finding People in Images and Videos," PhD Thesis, Institut National Polytechnique De Grenoble, 2006.

[8] N. D. Thanh, W. Li and P. Ogunbona, “A Novel Template Matching Method for Human Detection,” in 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, pp. 2549-2552.2009

[9] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, “Actions as Space-Time Shapes,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Volume 29, Issue 12, p. 2247–2253, 2007.

[10] C. J. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, Volume 2, Issue 2, p. 121–167, Jun. 1998.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, “Distance transforms of sampled functions,” Tech.Rep., Cornell Computing and Information Science, 2004.