Combining Template Matching and Model Fitting for Human Body Segmentation and Tracking with Applications to Sports Training

Hao-Jie Li[1,2], Shou-Xun Lin[1], and Yong-Dong Zhang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
100080, Beijing, China
[2] Graduate School of the Chinese Academy of Sciences,
100039, Beijing, China
{hjli, sxlin, zhyd}@ict.ac.cn

Abstract. This paper present a method for extracting and automatic tracking of human body using template matching and human body model fitting for specific activity. The method includes training and testing stages. For training, the body shapes are manually segmented from image sequences as templates and are clustered. The 2D joint locations of each cluster center are labeled and the dynamical models of the templates are learned. For testing, a "seed" frame is first selected from the sequence according to the reliability of motion segmentation and several most matched templates to it are obtained. Then, a template tracking process within a probabilistic framework integrating the learnt dynamical model is started forwards and afterwards until the entire sequence is matched. Thirdly, a articulated 2D human body model is initialized from the matched template and then iteratively fit to the image features. Thus, the human body segmentation results and 2D body joints are got. Experiments are performed on broadcasted diving sequences and promising results are obtained. We also demonstrate two applications of the proposed method for sports training.

1 Introduction

Video based human activity analysis has attracted increasing attention due to the wide range of potential applications, such as rehabilitation, security and sports training [1]. Segmentation human body from video sequences and automatic tracking body joint locations are important preceding techniques. An extensive amount of work have been done in these areas and promising results have been obtained [1, 2]. However, most of prior works either rely on background subtraction for static scene or motion cues for dynamic scene to separate foreground objects. These schemes work poorly when background is unavailable and objects keep still for a relative long time in a dynamic scene. Also, current model based tracking systems usually need manual initialization of the model parameters. In this paper, we present an approach to extracting and automatic tracking human body using template matching and human body model fitting for specific activity in dynamic environment. The aim is to segment human body shape and obtain main body joint locations for sports training.

A. Campilho and M. Kamel (Eds.): ICIAR 2006, LNCS 4142, pp. 823 – 831, 2006.

Background subtraction, temporal differencing and optical flow are three typical techniques for motion segmentation [2]. Background subtraction approach detects foreground regions by differencing between current image and a reference background image which is suitable to relatively static background [3]. Temporal differencing approach extracts moving regions by differencing between two or three consecutive frames, which usually generates holes inside moving objects [4]. Optical flow methods can be used to detect moving objects in dynamic background but are sensitive to noise [5]. As for the specific kind of object, human body, some prior knowledge can be used. For example, Gavrila [6] proposed a “chamfer” system to detect pedestrians from a moving vehicle using coarse-to-fine template matching technique. However, this system didn’t exploit temporal coherence between frames so it needed repeating the coarse-to-fine matching frame by frame which is time consuming.

2D/3D model based human body tracking is the dominant approach for body pose recovery. In [7], Ju *et al.* proposed 2D cardboard model to represent human limbs as a set of connected planar patches and each patch can undergo 2D affine transform. By performing these transforms the model is fitted to image. Human body can also be represented as conic sections [8] or super-quadrics for 3D pose recovery. The tracking is implemented in the fashion of “predict-match-update” and the previous or initial model state is needed to predict the current state. Most of these works rely on manually initialization of the first frame before tracking.

Our proposed method combines template matching and model fitting to segment human body and get its pose parameters. By template-matching to the selected “seed” frame and matching to neighboring frames in a fashion of tracking using the learnt dynamical model, the body shape of each frame is extracted and the initial guess of body model parameters are also obtained. Then the model is fitted by searching for the most likely pose parameters where the projection of the model is most similar to the appearance of real human in the image.

The outline of paper is as follows. The template hierarchy generating and dynamical model learning is described in section 2. In section 3, the selection and matching of the “seed” frame is discussed. The template tracking is presented in section 4. In section 5, the model fitting is described. Experimental results and applications are given in section 6. We conclude in section 7.

2 Template Hierarchy and Dynamical Model

In the training stage, we first cluster the manually segmented body shapes (i.e. templates) into a hierarchy, which facilitates a coarse to fine search strategy with a large amount of templates. Using the seven Hu’s moments of templates, a hierarchical *k*-means clustering algorithm is applied. The non-leaf nodes of the tree are the cluster centers of respective level and the leaf nodes are composed of templates themselves. Through a GUI the main body joints of each cluster center are labeled.

For specific activity, temporal coherence plays important role when tracking [9]. In our method, after matching the “seed” frame by searching the template tree, a tracking process using dynamical model is adopted to match the neighboring frames for its efficiency. The transition probabilities of clusters in the parent level of leaves are learnt. The transition probability from cluster *i* to *j* is estimated by:

$$p(c_j \mid c_i) = \frac{\sum_{t_p \in c_i} \sum_{t_q \in c_j} v(t_p t_q)}{\| c_i \|} \tag{1}$$

where $\| c_i \|$ is the number of templates in cluster c_i , $v(t_p t_q)$ =1 if template pair (t_p, t_q) appears in training sequences otherwise 0. In order to track in backward direction, the inverse transition probabilities are also learned.

3 Selection and Matching of the "Seed" Frame

Given the test sequence, we need not necessarily match the initial frame. Instead, we select a "seed" frame. The aim is to exploit more cues such as foreground area, not just the object contour, to make the matching more robust and efficient. It is clear that when matching using object contour only, it would produce many false positives[6] and a exhaustive search of the image is needed.

To select a "seed" frame, we first apply our spatial-temporal segmentation algorithm [10] to the entire sequence. As other motion segmentation algorithms, this algorithm generates good results when object have salient motion between successive frames while works poorly when objects have small motion (see Fig.1). The criterion for selecting "seed" frame is the reliability of motion segmentation. To measure the reliability, we define $\{s_i\}$ as the segmented foreground size for the sequence and μ the mean size. First, we select a sub-sequence starting from frame *m* with length *l,* which has the minimum foreground size variance and its mean size is larger than μ **T. T* is a predefined threshold. That is:

$$m = \arg\min_i \sum_{j=i}^{i+l} (s_j - \mu) * (s_j - \mu), \text{ subject to } \mu_m > \mu * T \tag{2}$$

where μ_m is the mean foreground size of sub-sequence *m*. Then we select a frame *k* from sub-sequence *m* as "seed" frame as follows:

$$k = \arg\min_i | s_i - \mu_m | \tag{3}$$

After the "seed" frame is selected, a top-down hierarchical template matching process is performed around the foreground area. The templates in the first level are matched at coarse grid locations. Only the sub-trees under templates having higher matching score are further explored. Thus the total number of template evaluations needed to search through the complete set is reduced.

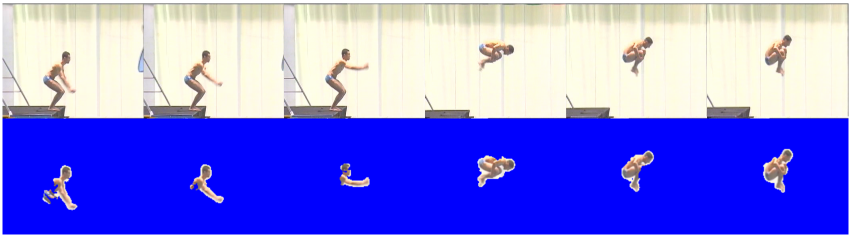

Fig. 1. Motion segmentation results for a diving clip. The images in upper row are the origin frames 25, 26, 27, 77, 78 and 79. The images in lower row are the segmentation results. In this example, frame 77 is selected as "seed" frame.

As stated above, each template is matched to image based on multi-cue, herein shape and foreground area. For shape, we adopt the distance transform [6] as similarity measurement between image edge features and template edge features. Let *d* be the mean sum of each edge's cost, the shape matching score for template t_i is:

$$p(z_{shape} \mid t_i) = \exp(-\lambda_1 * d) \tag{4}$$

For foreground, we compute the template's overlapping rate *r* to the foreground area, and then the foreground matching score is:

$$p(z_{foreground} \mid t_i) = \exp(\lambda_2 * r) \tag{5}$$

The total matching score is:

$$p(z \mid t_i) = p(z_{shape} \mid t_i) * p(z_{foreground} \mid t_i) \tag{6}$$

To make more robust, several most matched templates for the "seed" frame are retained and each retained template is transformed by sizing and rotating to get finer matching results. The template with the largest score is used to construct color histogram of the target which is used in the tracking process.

4 Template Tracking Within a Probabilistic Framework

For the rest frames of the sequence, we can repeat the hierarchical template matching to get the body shape. However, the searching is still time consuming even in the hierarchical manner. Due to the continuity of motion, the variation of pose, position and size of human body in neighboring frames is smooth and for specific activity the variation complies with some dynamical mode [9]. So the state of next frame can be predicted from previous known state and learnt dynamical model. Here we use a probabilistic framework, i.e. particle filtering as it is a popular approach used in visual tracking and provides an efficient way to approximate non-linear, non-Gaussian posterior distribution[11].

The state parameters needed to be estimated is the pose, position and size of human body corresponding to a state vector $X = (m, x, y, \theta, s)$. Here, m is the template index in library which includes the posture information and (x, y, θ, s) indicate the global position, rotation and size factor of template m respectively. In practice, we decompose the state vector into $X_1 = (m)$ and $X_2 = (x, y, \theta, s)$. For a particle, $X_1 = (m)$ is first sampled from the dynamical model and $X_2 = (x, y, \theta, s)$ is then sampled from a Gaussian distribution. That is

$$p(X_t \mid X_{t-1}) = p(X_{t,1}, X_{t,2} \mid X_{t-1,1}, X_{t-1,2}) = p(X_{t,1} \mid X_{t-1,1}) * p(X_{t,2} \mid X_{t-1,2}) \quad (7)$$

$$P(X_{t,2} \mid X_{t-1,2}) = N(X_{t-1,2}, \sigma_2) \quad (8)$$

The sampling of $X_{t,1}$ is divided into two steps. First the cluster c_t is sampled by $c_t \sim p(c_t \mid c_{t-1}(X_{t-1,1}))$ and the template index is then sampled from the cluster c_t with uniform distribution where $c_{t-1}(X_{t-1,1})$ indicates the cluster index of template $X_{t-1,1}$ and $p(c_t \mid c_{t-1}(X_{t-1,1}))$ is the learnt transition probability in Section 2. The filtering algorithm is outlined as follows:

```
1. Initialization
Draw a set of particles from the prior $p(X_{seed})$ to obtain $\{(X_t^{(i)}, w_i)\}_{i=1...N}$, t=1
2. Sampling step
    2.1 for i=1...N, sampling $X_t^{(i)}$ in two steps as above.
    2.2 evaluate the weights: $w_t^{(i)} = p(z_t \mid X_t^{(i)})$, i=1...N
    2.3 normalize the weights: $w_t^{(i)} = w_t^{(i)} / \sum_{j=1}^{N} w_t^{(j)}$, i=1...N
3. Output step
    The particle with the maximal weight is selected the output.
4. Resampling step
    Resample particles $X_t^{(i)}$ with probability $w_t^{(i)}$ to obtain N i.i.d random
    particles $X_t^{(j)}$ and set $w_t^{(i)}$=1/N, i=1...N
5. t=t+1, go to step 2.
```

Note that in step 2.2 when computing the likelihood of each particle, we consider one additional image cue, i.e. color histogram which is obtained from the "seed" frame. Since the "seed" frame is not necessary the initial frame of the sequence, the tracking should be made in two directions of the "seed" frame.

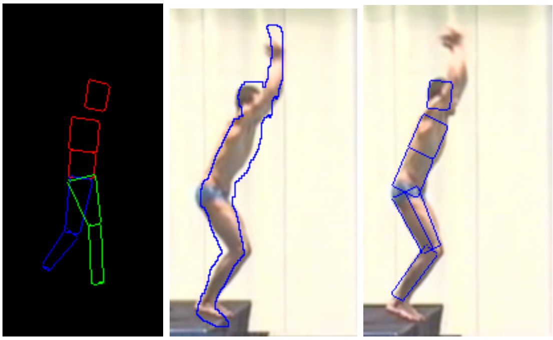

Fig. 2. Human body model and an example of model fitting result

5 Articulated Model Fitting

Fitting an articulated human body model to image cues is the popular approach to obtaining the posture information such as body joint angles or joint locations [1]. In our case, we adopt a 2D cardboard model [7] which is composed of 10 DOFs(degree of freedom), including 6 body joint angles and 4 global DOFs for position, rotation and sizing parameters. The template matching and tracking in Section 3 and 4 provide the initial guess of model parameters. The parameter vector is decomposed and a hierarchical search is applied near the initial state. The torso and thighs are first fitted, and then the head and calves.

Like template matching, the model's fitting score also comprises three parts: edge, foreground and color histogram. The human body model and a example of fitting are shown in Fig.2.

6 Experimental Results and Applications

6.1 Experimental Results

To test the validity of proposed method, we collect 28 clips of broadcasted diving video of 10m Platform. These clips includes 4 actions and each action has 6~8 clips. Two third of the clips are used for training and the rest for testing. We manually segment and label the templates, cluster them into a hierarchy and learn the transition probabilities of template pair.

6.1.1 Human Body Segmentation

Motion segmentation algorithm [10] is first used to segment the entire test sequence to select the "seed" frame. Some sample frames and corresponding segmentation results are shown in Fig.1. It is obvious that motion segmentation algorithm depends heavily upon the motion amplitude between successive frames. After template-tracking with learned dynamical model, some tracked results and segmented foreground objects are listed in Fig.3.

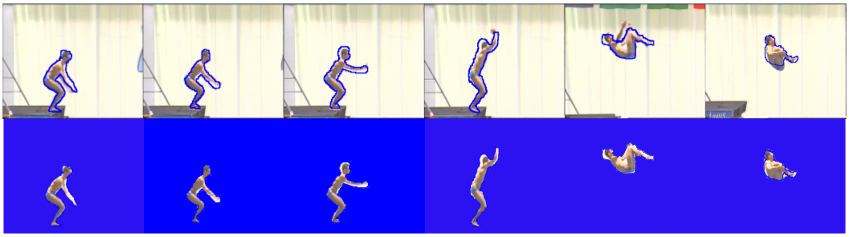

Fig. 3. The template tracking and segmentation results of proposed method for a diving clip (corresponding to frame 25, 26, 27, 33, 57 and 86). Compared to the segmentation results of frame 25-27 in Fig.1, the proposed method achieves better results when human have little motion.

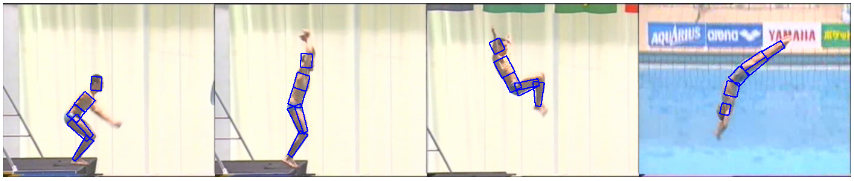

Fig. 4. Model fitting results for the diving clip

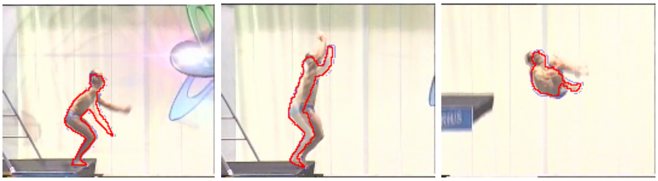

Fig. 5. Contrastive frames of same action performed by two athletes from two clips

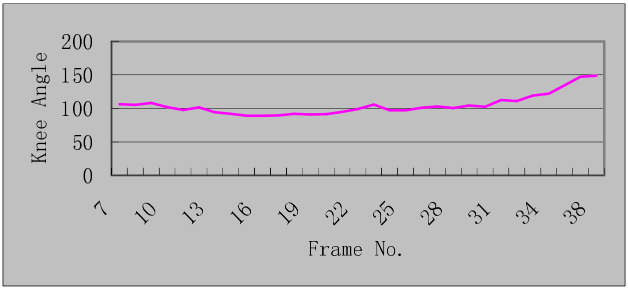

Fig. 6. Automatic obtained knee joint angle of the diver during take-off period

6.1.2 Human Body Model Fitting

After the most matched template to each frame is obtained, the state parameters of human body model is automatically initialized from the cluster center which the most matched template belongs to. Then a local search is performed to decide the final state. Some experimental results are shown in Fig.4.

6.2 Applications for Sports Training

In the field of sports training, visual cues analysis is becoming an important tool to improve athletes' performance. Here we demonstrate two applications of our method.

Given a synchronization point of two clips and the segmentation results, we can compose a contrastive clip by superimposing one clip's foreground on the other clip using alpha-composition technique. In the contrastive clip, nuance between actions can be easily found. Fig.5 demonstrates some example frames of a contrastive clip.

Biometric information is also useful for the coaches to instruct the athletes more scientifically. For diving, the knee joint angle during take-off period is critical since it decides the height of diving and influences the quality of the entire action directly. Here we apply our method to get the knee joint angle automatically as Fig.6.

7 Conclusions

In this paper, we present an approach to extracting and automatic tracking human body using template matching and human body model fitting for specific activity. It is particularly useful for sports training. Though this approach has the limitation that it need large number of training samples to build the template tree for accurate tracking, it is general and can be extended to other kind of activities such as trampoline and gymnastics.

Acknowledgements

This work is supported by the National Nature Science Foundation of China (60473002), the Key Project of International Science and Technology Cooperation (2005DFA11060), the Key Project of Beijing Natural Science Foundation (4051004), and Beijing Science and Technology Planning Program of China (Z0004024040231).

References

1. TB Moeslund and E. Granum. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding ,81(3): 231-268,2001
2. L Wang, W Hu and T Tan. Recent developments in human motion analysis. Pattern Recognition, 36 (3): 585-601, 2003
3. C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. Proc. of CVPR, pp. 246-252, 1999
4. A.J. Lipton, H. Fujiyoshi and R. S. Patil. Moving target classification and tracking from real-time video. Proc. of WACV, pp. 8-14,1998
5. D. Meyer, J. Denzler and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis. Proc. of ICIP, pp. 78-81,1997

6. D. M. Gavrila and V. Philomin. Real-time Object Detection for Smart Vehicles. Proc. of ICCV, pp. 87-93, 1999
7. S. Ju, M. Black and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. Proc. of AFGR, pp. 38-44 ,1996
8. J. Deutscher, A. Blake, *et al*. Articulated body motion capture by annealed particle filtering. Proc. of CVPR, pp. 1144-1149, 2000
9. H.Sidenbladh, M. Black and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. Proc. of ECCV, pp. 702-718,2000
10. Wu Si, Yong-Dong Zhang and Shou-Xun Lin. An automatic segmentation algorithm for moving objects in video sequences under multi-constraints. Proc of ICME, pp.555-558, 2004
11. M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. Proc. ECCV, pp. 343-356, 1996