A New Approach to Human Motion Sequence Recognition with Application to Diving Actions

Shiming Xiang[1], Changshui Zhang[1], Xiaoping Chen[2], and Naijiang Lu[3]

[1] State Key Laboratory of Intelligent Technology and Systems,
Department of Automation, Tsinghua University, Beijing 100080, China
{xsm, zcs}@mail.tsinghau.edu.cn
[2] Department of Physical Education, Tsinghua University, Beijing 100080, China
xpzhq@tsinghua.edu.cn
[3] Shanghai Cogent Biometrics Identification Technology Co. Ltd., China
lunj@cbitech.com

Abstract. Human motion sequence-oriented spatio-temporal pattern analysis is a new problem in pattern recognition. This paper proposes an approach to human motion sequence recognition based on 2D spatio-temporal shape analysis, which is used to identify diving actions. The approach consists of the following main steps. For each image sequence involving human in diving, a simple exemplar-based contour tracking approach is first used to obtain a 2D contour sequence, which is further converted to an associated temporal sequence of shape features. The shape features are the eigenspace-transformed shape contexts and the curvature information. Then, the dissimilarity between two contour sequences is evaluated by fusing (1) the dissimilarity between the associated feature sequences, which is calculated by the Dynamic Time Warping (DTW), and (2) the difference between the pairwise global motion characteristics. Finally, sequence recognition is performed according to a minimum-distance criterion. Experimental results show that high correct recognition ratio can be achieved.

1 Introduction

The recent years have seen a surge of interest in video-based human action recognition [1][2][3][4] . However, due to the non-rigidity of human body, human motion classification is a challenging problem. The key difficulty of classification is how to derive the time-varying information from image sequences for action segmentation [4] and motion sequence recognition [3]. Most works [2][4][5] have been done on partitioning an image sequence involving human into key frames, meta-actions, or meta-gestures for video content analysis, human computer interaction, virtual reality, behavior understanding, or sign language recognition. Instead of aiming at analyzing the details within a single image sequence, comparing between different image sequences is desired for intelligent surveillance, content-based video retrieval, video-assisted analysis in athletic training and heath-care arenas, and entertainment, etc..

To obtain the motion information from an image sequence, the motion detection and human tracking methods [1][2] can be employed to obtain a sequence of binary silhouettes or a sequence of pose parameters. Since this sequence is associated with the human body, it can reflect the spatio-temporal motion information. We can then derive time-varying feature sequences [3][6] or calculate some important motion properties, such as speed, period, amplitude, number of somersaults in diving, etc.. For example, gait recognition [3] aims to signify the identification of individuals in the image sequences by their gait styles. However, in many applications, identifying who is in an image sequence may be unnecessary. Instead, identifying the motion type to which the motion belongs is desired.

Since gait is a biometric feature, the methods [6] to be used to extract gait feature sequences may not be directly applied to other motions, such as jumping, diving, etc.. Sequence feature analysis for these situations is a new problem.

This paper aims to identify the action group to which the dive belongs. To this end, each image sequence is converted to a 2D contour sequence by our exemplar-based tracking approach. The reasons we analyze 2D contour sequences are: (1) The deformations of the contour can reflect the changes of the pose configuration; (2) Shapes are more robust to the changes of clothing and illumination than color and texture.

The recognition strategy is constructed for the whole 2D contour sequences. We use eigenspace-transformed shape contexts [7] and curvature information as shape features. The features of all contours are listed over time to form a feature sequence. Fig. 1 illustrates the process.

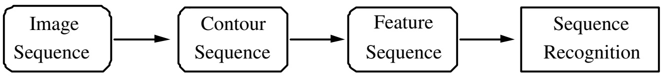


Fig. 1. The process of sequence recognition

Besides the feature sequence, we also use the number of somersaults, which is one of the most distinct global motion characteristics in diving, to describe the 2D contour sequence as a whole. The dissimilarity between two feature sequences is computed by sequence matching through Dynamic Time Warping (DTW) [8] approach. To decide the final dissimilarity between two contour sequences, the dissimilarity of two feature sequences and the difference of global pairwise characteristics are integrated together. Finally, sequence recognition is performed according to a minimum-distance criterion (see Fig. 1).

This paper is structured as follows. Section 2 briefly introduces the related work. Section 3 details the proposed simple and effective approach to deformable contour tracking. The feature analysis approaches to contour sequence are described in Section 4. Section 5 outlines the algorithm of sequence recognition. The experimental results are reported in Section 6, followed the conclusion in Section 7.

2 Related Work

This section briefly reviews the related work on shape representation and motion sequence analysis. The literatures on shape representation are rich [9]. However, we do not need those representations with rotation invariant features, such as Hu moments, Fourier descriptors, and those wavelet based features (see [9] for details), because diving motion is highly related to the rotation of the human body. Whereas, most rotation sensitive representations can only capture the global perception characteristics, for example the spatial moments [9], and hence are incapable of describing the local shape feature well. Belongie et al. proposed a novel method for shape representation and shape matching [7]. The basic idea of their proposal is to construct a shape context for every discretized contour point. Due to the detailed description, measuring the similarity between two points from two shapes can be done explicitly.

From the point view of pattern recognition, two basic tasks are related to image sequence analysis. One task is to partition a sequence into different meta-poses or meta-actions [4]. The other task is to recognize image sequences based on a sequence gallery by taking each of them as a probe sequence. Each probe sequence is described by global motion characteristics or converted to an associated feature sequence. The global characteristics can be derived from time-independent features [10] or time-related features [11]. A feature sequence is a temporal sequence of features, such as the sequences derived from gait styles [3][10][12]. However, different kinds of motions have their own characteristics. Thus, extracting the salient feature is crucial for sequence recognition.

Sequence recognition is performed according to feature comparison. Currently, most of the related works are developed for gait recognition [3][6][10][12] . In contrast, the Hidden Markov Model (HMM) based methods [4][12] and the DTW [8] based methods [3] are more suitable for general sequence comaparison. To use HMM, it is necessary to partition the sequences into meta-actions, meta-gestures or key frames as samples to learn the model parameters. The DTW is a common technique since there is no need for one to learn the prior model. However, we need to prepare the sequences to be recognized with roughly equal sequence lengthes, according to the work of Rabiner et al. [8].

3 Contour Extraction

We use target tracking approach to extract the contours since the background is non-static. For visual-based human tracking [2], Sequential Monte Carlo (SMC) estimation [13] has proved to be a successful approach. In SMC framework, the probability of the object configuration given the observation is described by a set of weighted particles. Tracking process can then be viewed as a density propagation governed by the dynamic model and observation model [14].

Dynamic model is highly related to contour representation. Due to non-rigid motion and occlusions during diving, representing the 2D deformable diver contours is a tough task. The efficient method with regards to processing deformation is to define complex model with high dimensionality. However, in SMC framework, this leads that the density function which governs the distributions of the target states would be propagated in a high-dimensional state space. It seems that we can use parameterized curves to describe the contour. But due to occlusions of arms, the changes of the 2D pose configuration are drastic.

However, we observe that in diving there exist fundamental poses, which can be used to depict the new ones. To this end, we collect the fundamental contours from different diving action groups to construct a database of exemplars (denoted by E). We use the exemplars to describe the appearances of the target states as well as guide the tracking process. As a result, we can only use three state variables in dynamic model, namely, the centroid coordinate (x, y) and the scale parameter s. Now we can write the dynamic equation as follow:

$$\begin{cases} x_t = x_{t-1} + V_t(x) \\ y_t = y_{t-1} + V_t(y) \end{cases} \tag{1}$$

where (x_t,y_t) is the centroid coordinate of the target state at time t, and $V_t(x)$ and $V_t(y)$ bear normal distribution $N(0, \sigma_x)$ and $N(0, \sigma_y)$, respectively.

Each particle (x_t,y_t) employs an exemplar as its appearance, scaling a little with parameter s. s is randomly set within the range of 0.9-1.1 since the camera was always located in the same place at a distance from the diving platform.

We use the exemplars approximately corresponding to the standing poses to initialize the particles' appearances. After scaled with s, each of them is located in the first frame by using fast Hausdorff distance mapping [15].

Then, we embed a process of contour recognition into the tracking process. For the associated contour of a particle, we retrieval its neighbors from E as its candidates, which are distributed in the current frame according to Equation 1, respectively. After measured through observation model [16], the one with the maximum posterior probability is selected and transferred to the next frame.

To fast retrieval the needed neighbors, the contours in E are organized as a tree structure, based on all the two and three order contour moments.

Toyama et al. perform probabilistic tracking with exemplars in a metric space [17]. There the exemplars are interpreted probabilistically. However, we use exemplars as inputs for searching to find the candidates. Furthermore, we do not need complex training process. Actually, neighbor search approach provides the updating dynamics for particles' appearances as well as the mechanism to guide the tracker to find the candidates for each particle. Our method is a simple and effective approach for the purpose of sequence recognition. Fig. 2 shows some tracked frames from three image sequences.

The exemplar database includes 210 different 2D contours. During tracking, the particle number is 4000 and the number of neighbors to be searched is 10. We manually take $\sigma_y = 2\sigma_x$ and $\sigma_x = 8$ since the motion of the centriod of the diver body is roughly controlled by gravity and the motion in the horizontal direction is limited.

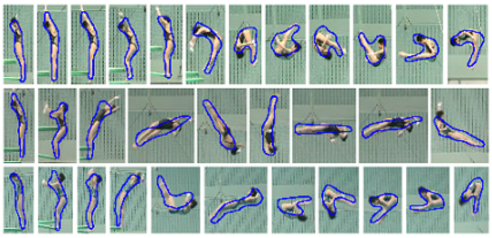


Fig. 2. Some results of contour tracking from three image sequences, respectively

4 Sequence Recognition

4.1 Feature Sequence

To convert a contour sequence into a feature sequence, we need the shape features with translation and scale invariance since the contours are translated to the image centers and the body sizes of the divers may be slightly different.

We use shape context descriptor as shape feature. For each reference point, its shape context is a log-polar histogram of the relative coordinates of the remaining points. The shape context summarizes global shape in a rich and local descriptor. Since each point can be associated with a histogram, we can get a shape context matrix, which is a detailed description about the shape perception.

Invariance to translation is intrinsic to the shape context. To achieve scale invariance, all radial distances by the median distance between all the point pairs is normalized [7].

We observe that for most shape contexts a lot of bin values are zeros. This results that the histograms are sparse. Directly using the χ^2 statistics to measure two sparse histograms may not reflect the similarity well [18]. Thus we apply the eigenspace transformation based on Principal Component Analysis (PCA) to the histograms to reduce the redundancy. The details are as follows:

We use all the shape contexts calculated from E as PCA training samples. After performing PCA, we take k eigenvectors corresponding to the k largest eigenvalues, $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}$, to form an eigenspace $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]$. For a novel histogram vector $\mathbf{X}$, we have:

$$\mathbf{Y} = \mathbf{E}^T \mathbf{X} \tag{2}$$

On the other hand, the log-polar space makes the shape descriptor more sensitive to the positions near the reference point. In fact, it is unable to robustly reflect the local geometrical property very well. Actually, the degree of curvature is highly related to a few neighbor points. We use it as an additional feature.

Now a contour is described as a group of features $\{\mathbf{C}, \mathbf{K}\}$, where $\mathbf{C}(\in R^{N\times k})$ is the eigenspace-transformed shape context matrix, and $\mathbf{K}(\in R^N)$ is the curvature vector. Here N is the number of discretized points of the contour and M is the number of the bins of the shape context histogram. As a result, a contour sequence is naturally converted into a feature sequence.

Let point P_S^i belong to contour S, and P_T^j belong to T, the distance between P_S^i and P_T^j can then be computed as follow:

$$d(P_S^i, P_T^j) = \chi^2(\mathbf{C}_S^i, \mathbf{C}_T^j) + s_1 \cdot d_{s2}(\mathbf{C}_S^i, \mathbf{C}_T^j) + s_2 \cdot d_k(\kappa_S^i, \kappa_T^j) + s_3 \cdot d_{k2}(\kappa_S^i, \kappa_T^j) \quad (3)$$

where $\mathbf{C}_S^i$ and κ_S^i denote the eigenspace-transformed shape context and the curvature of the i^{th} point of contour S. $\mathbf{C}_T^j$ and κ_T^j have the same meanings as $\mathbf{C}_S^i$ and κ_S^i, respectviely. s_1, s_2 and s_3 are weighting parameters, which are all manually set as 0.001.

In Formula 3, $\chi^2(P_S^i, P_T^j)$ and $d_{s2}(P_S^i, P_T^j)$ are calculated as the χ^2 statistics and the two order derivative of the eigenspace-transformed shape context cost at the pair point of (P_S^i, P_T^j) [19][20]. $d_k(P_S^i, P_T^j)$ and $d_{k2}(P_S^i, P_T^j)$ are the curvature cost and the two order derivative of the curvature cost, respectively. The reason here we use the two order derivatives is that close points on S should also be close after matched to T.

Finally, the similarity between S and T can be determined, by performing shape matching [7] based on Formula 3.

4.2 Global Motion Characteristics

The number of somersaults (denoted by $\varPi$) is a salient global motion characteristic. To calculate $\varPi$, we track the position of the feet to form a trajectory and then calculate the rotation number. It is feasible since during diving the feet are always keeping straight and close together and seldom occluded by the arms in the sky. This leads the diver contours have thin appearances. Thus we can extract their skeletons. Now the steps to calculate $\varPi$ can be summarized as follows:

First, extract the skeleton by morphological thinning operation and trim off the branches with small lengths. Then, detect the branch ends and track the one corresponding to the feet based on the movement continuity. To perform this step, all the vectors defined from the image center to the ends are first normalized. The vector with minimum angle to the tracked vector in the previous frame is selected as current result. Thus, we get a normalized trajectory. Due to translation and normalization, it does not correspond to the real physical one. However, this does not affect the calculation because the diver body turns approximately along its own axis. Finally, $\varPi$ can be computed as:

$$\varPi = \frac{1}{2\pi} \sum_{i=1}^{N} sgn(\mathbf{v}_{i-1} \cdot \mathbf{v}_i) \cdot \arccos(\mathbf{v}_{i-1} \cdot \mathbf{v}_i) \quad (4)$$

where $sgn(\mathbf{v}_{i-1} \cdot \mathbf{v}_i)$ stands for the relative rotation direction from $\mathbf{v}_{i-1}$ to $\mathbf{v}_i$, and N the is total frame number. $sgn(\mathbf{v}_{i-1} \cdot \mathbf{v}_i) = 1$ means the rotation direction is counter-clockwise, while $sgn(\mathbf{v}_{i-1} \cdot \mathbf{v}_i) = -1$ means the rotation direction is clockwise. Here, $\mathbf{v}_0$ is the initial position vector.

5 Sequence Recognition Algorithm

Note that the lengths of the image sequences would be very different. To make the sequences to be recognized with almost equal lengthes for using DTW matching, we cut off the frames corresponding to the preparing stage because the poses are rest stances and hence weakly informative in the context of action recognition.

To this end, we use again the normalized trajectory. By finding the point which begins to depart from the vertical position, we obtain the corresponding image frame. Thus the sequence can be partitioned into two subsequences. We take the later one for recognition.

The contour sequences may be different every time since the divers may slightly adjust their poses and alter or control the motion speed. Directly performing frame-to-frame matching is not realistic. Therefore, We use the DTW to match the sequences and define the matching cost as dissimilarity [3][8].

Let $S_1 : \{S_1^1, \cdots, S_1^n\}$ and $S_2 : \{S_2^1, \cdots, S_2^m\}$ be two contour sequences. Let $\mathbf{C}_i^j$ and $\mathbf{K}_i^j$ denote the eigenspace-transformed shape context matrix and the curvature vector of S_i^j, respectively. Suppose the number of somersaults of S_i be Π_i. We summarize the steps of computing the dissimilarity between S_1 and S_2 as follows:

Step1: Calculate $\mathbf{C}_1^j$, $\mathbf{K}_1^j$ $(j = 1, \cdots, n)$ and $\mathbf{C}_2^j$, $\mathbf{K}_2^j$ $(j = 1, \cdots, m)$;

Step2: Transform $\mathbf{C}_1^j$ $(j = 1, \cdots, n)$ and $\mathbf{C}_2^j$ $(j = 1, \cdots, m)$, according to Formula 2;

Step3: Calculate the distance matrix $\mathbf{M} \in (R^{n \times m})$ for S_1 and S_2:

(1) for S_1^i and S_2^j $(i = 1, \cdots, n;\ j = 1, \cdots, m)$, compute the pairwise matching cost $e_{i,j}$ based on Formula 3,

(2) let $M_{i,j} = e_{i,j}$;

Step4: Based on $\mathbf{M}$, use the DTW matching to calculate the matching cost, and denote it by d_1;

Step5: Calculate Π_1 and Π_2, and let $d_2 = |\Pi_1 - \Pi_2|$;

Step6: Compute the dissimilarity between S_1 and S_2: $d = \sqrt{d_1^2 + (wd_2)^2}$.

Finally, a probe sequence is identified based on the minimum-distance criterion.

6 Experimental Evaluation

The raw video data involving the divers in training was taken at a distance by a CCD camera in different days. To keep the diver figures in the range of the image plane, the camera may slightly rotate along the camera support.

We use the second and fourth group of the international standard diving actions to test our method. The second group action, denoted by ‘2’, is the back group (face to the platform or springboard at the beginning, diving backward). The forth group (‘4’) is the inward group (diving inward). There have four fundamental pose groups, denoted by ‘A’ (straight), ‘B’ (pike), ‘C’ (tuck) and ‘D’ (free), respectively. The parameters about somersaults are complex. ‘1’ stands for ‘0.5’ number of somersaults, ‘2’ for ‘1.0’ number of somersaults, etc. Thus, according to international diving criteria, '21A' means "the second group, 0.5 number of somersaults, straight pose".

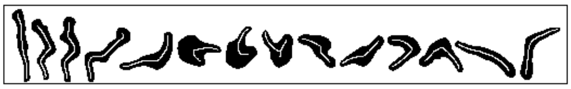
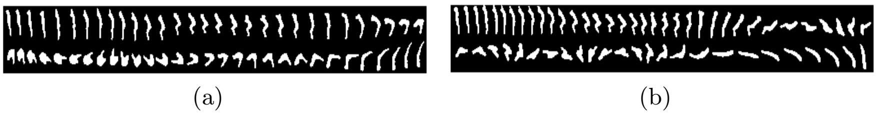


Fig. 3. Some 2D poses and their skeletons extracted by morphological thining

We build a gallery including of 10 groups of diving actions from a single diver: 21A, 23A, 23B, 25B, 23C, 25C, 21D, 23D, 43B, 43C. All the ten image sequences are converted into ten 2D contour sequences by hand. Then a database consisting of 210 different poses is constructed by selecting the same pose once. It is used for both contour tracking (see Sect. 3) and PCA training.

The size of the shape context histogram is 5×12. The number of the discretized contour points is 80. Thus, there are totally 16800 samples for PCA training. When taking the eigenvectors, we let $k = 20$.



(a) (b)

Fig. 4. Two 2D contour sequences extracted by using the method in Sect. 3

Fig. 3 shows the 2D poses and their branch-trimmed skeletons. Fig. 4 gives two translated sequences. The diving code in Fig. 4(a) is '43B', while the code in Fig. 4(b) is '25C'. Fig. 5 demonstrates the translated and normalized trajectories. The trajectories demonstrated in Fig. 5(a) and Fig. 5(b) correspond to the sequences in Fig. 4(a) and Fig. 4(b), respectively. We can see that two sequences have the same length. But they show different pose configuration and different rotation direction (See in Fig. 5). The real values of the numbers of somersaults of the diving actions shown in Fig. 4(a) and Fig. 4(b) are 1.5 and -2.5, respectively. The corresponding values calculated from Formula 4 are 1.45 and -2.38, respectively.

The testing set includes 50 image sequences involving four divers. Each image sequence is used as a probe sequence. The task is to recognize the gallery sequence corresponding to the probe sequence. We use contour tracking to obtain a 2D contour sequence for a probe sequence. The action group of the probe sequence is identified as that of the gallery sequence with which the matching distance is minimum, according to the algorithm in Sect. 5. We achieve 100% correct recognition ratio for 50 testing sequence.

Fig. 5. The translated and normalized trajectories

7 Conclusion

This paper aims to recognize diving actions directly based on image sequences. Different from the traditional work on action recognition, we treat the sequence as a whole, rather than partition it into different meta-actions or key frames. We use exemplar-based contour tracking to convert an image sequence into a 2D contour sequence. The eigenspace-transformed shape context histogram matrix and curvature information are used as shape features to form a feature sequence. The dissimilarity of two feature sequences is determined by sequence matching.

The global motion characteristics and motion type determined by the recognition framework of this paper are important video contents for content-based video retrieval and video mining. The meta-data based methods can only summarize the global perception information, which is produced jointly by the humans and the other uninteresting objects.

Although the work is developed on diving actions, the proposed approaches to visual tracking, sequence feature analysis may be applied to other visual computations or video analysis tasks since the related problems are general. In the future, experiments on bigger database and more actions will be carried out to test our method. And we would like to develop more general method for human motion sequences-oriented spatio-temporal pattern analysis.

Acknowledgements

This work is supported by the project (60475001) of the National Natural Science Foundation of China

References

1. Gavrila,D.: The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding, **73** (1999) 82-98
2. Wang, L., Hu, W.M., Tan, T.N.: Recent Developments in Human Motion Analysis. Pattern Recognition, **36** (2003) 585-601
3. Wang, L., Tan, T.N., Ning, H.Z., Hu, W.M.: Silhouette Analysis-based Gait Recognition for Human Identification. Transactions on Pattern Analysis and Machine Intelligence, **25** (2003) 1505-1518
4. Wu, Y., Huang, T.: Vision-based Gesture Recognition: A Review. In: Proceedings of the International Gesture Workshop, Gif-sur-Yvette France (1999) 103-115

5. Cedras, C., Shah, M.: Motion-based Recognition: A Survey. Image and Vision Computing, **13** (1995) 129-155
6. Nixon, M.S., Carter, J.N.: Advances in Automatic Gait Recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, Seoul Korea (2004) 139-144
7. Belongie,S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence. **24** (2002) 509-522
8. Rabiner. L., Juang, H.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey (1993)
9. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition, **31** (1998) 983-1001
10. Lee,L., Grimson, W.E.L.: Gait Appearance for Recognition. In: Proceedings of European Conference on Computer Vision, Copenhagen Denmark (2002) 143-154
11. Collins,R.T.,Gross, R., Shi, J.B.: Silhouette-Based Human Identification from Body Shape and Gait. Proceedings of International Conference of Automatic Face and Gesture Recognition, Washinton D.C. USA (2002) 351-356
12. Kale, A., Sundaresan,A., Rajagopalan,A. N., et al.: Identification of Humans Using Gait. IEEE Transactions on Image Processing, **13** (2004) 1163-1173
13. Doucet, A., De Freitas, N., Ngordon, N.: Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York (2001)
14. Isard, M., Blake, A.: Condensation Conditional- Density Propagation for Visual Tracking. International Journal of Computer Vision, **26** (1998) 5-28
15. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.A.: Comparing Images Using the Hausdorff Distance. IEEE Transactions on Pattern Analysis and Machine Intellegent, **15** (1993) 850-863
16. Shen, C.H., van den Hengel, A., Dick, A.: Probabilistic Multiple Cue Integration for Particle Filter Based Tracking. In: Proceedings of Digital Image Computing: Techniques and Applications, Sydney Australia (2003) 399-408
17. Toyama, K., Blake, A.: Probabilistic Tracking with Examplars in a metric Space. International Journal of Computer Vision, **48** (2002) 9-19
18. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision, **40** (2000) 99-121
19. Thayananthan, A., Stenger, B., Torr, P. H. S., Cipolla, R.: Shape Context and Chamfer Matching in Cluttered Scenes. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, Madison Wisconsin (2003) 127-133
20. Srisuk,S., Tamsri, M., Fooprateepsiri, R., Sookavatana, P. Sunat, K.: A New Shape Matching Measure for Nonlinear Distorted Object Recognition. In: Proceedings of Digital Image Computing: Techniques and Applications, Sydney Australia (2003) 339-348