On the use of Radon Transform for Facial Expression Recognition

Nicolas Tsapatsoulis, Yannis Avrithis and Stefanos Kollias

Department of Electrical and Computer Engineering
National Technical University of Athens
Heroon Polytechniou 9, 157 73 Zographou, Greece
e-mail: ntsap@image.ntua.gr

ABSTRACT

A facial expression recognition scheme is presented in this paper, based on features derived from the optical flow between two instances of a face in the same emotional state. A pre-processing step of isolating the human face from the background is first employed by means of face detection and registration. A spatio-temporal description of the expression is then obtained by evaluating the Radon transform of the motion vectors between the face in its neutral condition and at the 'apex' of the expression. A linear curve normalization scheme is proposed, achieving a translation, scaling and resolution invariant representation of the Radon curves. Finally, experimental results are presented, illustrating the performance of the proposed algorithm for expression classification using a correlation criterion and a neural network classifier.

Keywords: Facial Expression Recognition, Radon Transform, Curve Normalization, Neural Networks.

1. INTRODUCTION

Two channels have been distinguished in human interaction [1]. One transmits explicit messages, which may be about anything or nothing: the other transmits implicit messages about the speakers themselves. Understanding the other party's emotions is one of the key tasks associated with the second, implicit channel. Building an emotion detection system makes it possible to assess how well ideas explain people's general competence at understanding emotion. Methods by which a computer can recognize visually communicated facial actions-expressions are of high importance since they can be used to categorize active and spontaneous facial expressions, so as to extract information about the underlying emotional states, using visual cues. Approaches to the recognition of facial expressions can be divided into two main categories: static and motion dependent. In static approaches, recognition of a facial expression is performed using a single image of a face. Motion dependent approaches extract temporal information by using at least two instances of a face in the same emotional state. When two instances are used (semi-static approaches), they usually represent the face in its neutral condition and the face at the peak ('apex') of the expression [2]. Fully dynamic approaches use several frames (generally more than two and less than 15) of video sequences containing a facial expression, which normally lasts 0.5 to 4 seconds [3].

In this paper we present an expression recognition scheme based on features derived from the optical flow between a neutral and a face at the 'apex' of the expression. We avoided using a fully dynamic approach due to the lack of large databases containing expression sequences. However, the method can be easily extended to cover the dynamic approach. The proposed algorithm utilizes Radon transform of the motion vectors amplitude to spatio-temporally describe the expressions. Since we concentrate on computing the basic perturbation within face area, extremely accurate estimation of the optical flow is not required [2][3]. Instead a pre-processing stage for face detection and normalization is necessary, to enable the computation of facial pixels movement and not that of the background ones. Optical flow is estimated only for those facial parts where substantial movement has occurred. Curves obtained from the Radon transform in different angles are used for classification. Finally a technique for curve normalization is included in the paper, and leads to simple classification either using correlation or a neural classifier.

2. FACE DETECTION AND REGISTRATION

As mentioned in the previous section we aim at extracting the facial pixels movement to proceed with expression description. A pre-processing stage where the face should be detected and registered is required.

Face Detection

Face detection, i.e., isolating the face from the background, is not a trivial task. However in expression recognition two constraints can be posed. First, there is only one face with significant scaling in the expression frames, thus no multiface searching is required, and second, the face is the same along the expression. According to the second constraint, aging and personal variations like wearing make-up, eyeglasses or beard, are not likely to occur during the expression. Furthermore,

once the face is detected in the first frame, gained information about its anatomy, scaling and orientation can be used in the subsequent frames to minimize the detection effort.

The algorithm for face detection is described next. Let $M(u,\theta)$ be a face template at scale *u(h,v),* described by horizontal scaling *h* and vertical scaling *v*, and angle (rotation) θ. Let *F* be a frame containing the face at arbitrary scale, location and rotation, and *A* a frame area with the same scaling and rotation as *M.* We use the following metric to find the minimum correlation between *A* and *M* at scale *u* and orientation θ:

$$r(u,\theta) = \min_{A \subset F} \left\{ \frac{|A - M(u,\theta)|}{\rho \cdot a \cdot b} \right\} \quad (1)$$

where $\rho = 1 - c \cdot \left| \frac{h}{v} - \frac{2}{3} \right|$, is used to account for the face anatomy, *c* is a constant (0<*c*<0.5) and $a = \text{mean}(A)$, $b = \text{mean}(M)$ are used to account for illumination variations.

The best scale and orientation are obtained by $[U,\Theta] = \arg\min\{r(u,\theta)\}$ and the final detection is performed using the template $M(U,\Theta)$ and the metric in (1). The corresponding detected area A^* is then scaled and rotated according to parameters *U,* Θ so that it is transformed to standard predefined co-ordinates.

The detection procedure described above is applied only to the first frame. In the frame of the expression's 'apex' detection is also performed but the best scale *U* and rotation Θ are known, thus computational complexity is significantly reduced.

Face Registration

Registration is the transformation of the face to standard predefined co-ordinates based on detecting the main facial features (eyes, nose and mouth) [4] and renormalizing the face by translating, rotating and expanding / shrinking it around a virtual central (nodal) point. To account for variations caused by facial expressions, an image warping transformation based on radial basis functions has been proposed which decomposes the transformations into linear and radial terms [5]. In our approach the face registration step is not included in the algorithm. Translation, rotation and scaling variations are removed in the face detection step, while expanding / shrinking is important only in face recognition tasks where one should account for different faces, and not in expression recognition where the face is the same during the expression.

3. OPTICAL FLOW ESTIMATION

In the following we estimate the optical flow directly from facial pixel values without involving facial region tracking or muscle modeling so as to avoid daunting analysis tasks such as edge detection and facial features (i.e., eyes, mouth, etc.) localization. The motion field is computed only in face areas where substantial movement has occurred [6]:

Let F_k and F_{k+1} be the neutral and 'apex' frames respectively, in which the face has already been detected and normalized. Each pixel $p_k(x,y)$ at the *k-th* frame is described through its surrounding $2n \times 2n$ block $b_k(x,y)$, and is associated with the following error:

$$e_k(x,y) = |b_k(x,y) - b_{k+1}(x,y)| = \\ = \sum_{l=-n}^{n} \sum_{m=-n}^{n} |p_k(x+l, y+m) - p_{k+1}(x+l, y+m)| \quad (2)$$

Motion vectors are calculated only for the blocks with significant $e_k(x,y)$ based on appropriate image-dependent thresholding. The motion vector $\hat{\mathbf{v}}_k(x,y) = (\hat{v}_x, \hat{v}_y)$ of block $b_k(x,y)$ is computed using block matching in a neighborhood of block $b_{k+1}(x,y)$ according to the equation:

$$\hat{\mathbf{v}}_k(x,y) = \arg\min_{(v_x,v_y) \in Q} \sum_{l=-n}^{n} \sum_{m=-n}^{n} |d_k(x+l, y+m; v_x, v_y)| \quad (3)$$

where

$$d_k(x,y;v_x,v_y) = p_k(x,y) - p_{k+1}(x - v_x, y - v_y) \quad (4)$$

and $Q = \{-q,...,q\} \times \{-q,...,q\}$ is the search area. In order to decrease execution time, logarithmic search is employed, i.e., only a limited subset of combinations $(\hat{v}_x, \hat{v}_y) \in Q$ are used for searching.

'Noisy' motion vectors (i.e., badly estimated pixel-motion) inevitably arise due to the simplicity of motion estimation; to account for this, median filtering is adopted. First, median filtering is applied to motion vectors phase (directional filtering) and then to their magnitude.

4. RADON TRANSFORM

Radon transform computed at different angles along with the estimated facial pixel motion gives a spatio-temporal representation of the expression. Facial areas, which move to produce the expression, are efficiently tackled.

The motion vector at position (*x*,*y*) can be expressed as $\hat{\mathbf{v}}_k(x,y) = a_k(x,y) e^{j\phi_k(x,y)}$. The discrete Radon

transform of the motion vector magnitude, at angle θ, is then given by:

$$R(\theta) = \sum_{u=-\infty}^{\infty} a_k(x,y)\Big|_{x=t\cos\theta - u\sin\theta,\, y=t\sin\theta + u\cos\theta} \quad (5)$$

As it can be seen from this equation, using the Radon transform we estimate the spatial distribution of the energy of perturbation. An attempt to compute the direction of the facial parts motion, although useful, requires an extremely accurate optical flow estimation as well as facial feature (eyes, mouth) tracking.

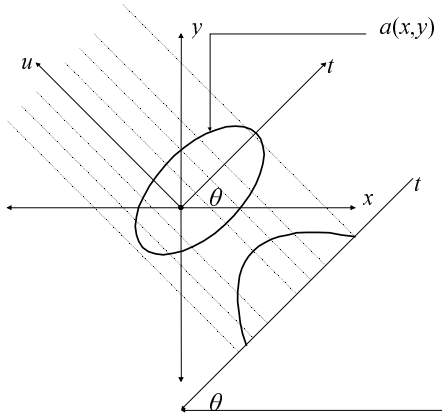

Figure 1. Illustration of the Radon transform.

The projections on Radon transform on two different angles, 0° and 90°, called 'signatures', are used to characterize the expressions. As it is described in the following section, signatures are normalized to standard co-ordinates, so that direct classification of signatures derived from images of different resolution or scaling is possible.

5. SIGNATURE NORMALIZATION

The normalization process is a transform invariant to image resolution, translation and scaling and consists of a set of linear operations on signatures. It is in fact a special (one-dimensional) case of a more general 2-dimensional affine-invariant curve transform that has been used for normalization and classification of curves representing object contours in image databases [7]. Signature normalization is necessary even when face registration is applied to the original image sequence, mainly in order to account for different image resolutions.

Let $R=[r_i]$, $i \in F = \{0,\ldots,L-1\}$ be the $1\times L$ vector of the Radon transform corresponding to angle 0° or 90°. Vertical scale normalization is performed first by calculating the vector $S=[s_i]$, $i \in F$, as

$$s_i = r_i \left(\frac{1}{L} \sum_{k=0}^{L-1} r_k^2 \right)^{-1/2}, \quad i = 0,\ldots,L-1 \quad (6)$$

Horizontal normalization follows next, effectively by removing zero values from the left and right edges of vector S. This step is necessary in order to retain the "central" portion of each image, that contains moving parts of a face. Defining $F' = \{i \in F : s_i > T\}$ as the set of indices of "non-zero" elements of S, and $i_L = \min\{i \mid i \in F'\}$, $i_R = \max\{i \mid i \in F'\}$ as the leftmost and rightmost elements of F', respectively, the horizontally normalized vector $Z=[z_i]$, $i=0,\ldots,i_R-i_L$ is derived as

$$z_i = s_{i+i_L}, \quad i = 0,\ldots,i_R - i_L \quad (7)$$

Threshold T is selected so that zero values of S are efficiently distinguished from non-zero ones. A value in the order of 0.1 is usually satisfactory, and this value is independent of signature values, since scale normalization has already been performed.

Finally, the normalized signature is a $1\times K$ vector $N=[n_i]$, $i=0,\ldots,K-1$ and is derived by resampling vector Z at K points, using linear interpolation. Thus, all normalized signatures are vectors of equal length and can be directly compared and classified. Since the Radon transform is computed at two different angles, 0° and 90°, a single vector of length $2K$ is constructed, containing the normalized signatures corresponding to these two angles. This vector is then used for classification using a correlation coefficient and a neural network approach, as described in the following section.

The normalization process, as defined above, has several important properties:

1. It is invariant to translation, scaling and number of elements (resolution) of signature vectors.
2. It consists of linear operations, so that no information is actually lost (apart from the zero elements, which are of no importance for signature classification).
3. The same transform is applied to all signatures, meaning that no extra knowledge is required for comparison or matching between two signatures. Normalized signatures can thus be directly used by any classification mechanism.

These properties are not found in other scale/translation-invariant techniques, such as Fourier descriptors and moments [8].

6. EXPERIMENTAL RESULTS

In the experiments we used 75 images of the Yale database which correspond to the expressions "normal", "happy", "surprised", "sad" and "sleepy". Normal

images were used as the first frame (neutral) of all other expressions. In Figure 2 four "normal" images are shown after the face detection procedure. Figure 3 illustrates the estimated motion and the facial areas with substantial movement for the expressions "happy", "surprised", "sad" and "sleepy".

Figure 2. Images of four subjects after face detection and registration.

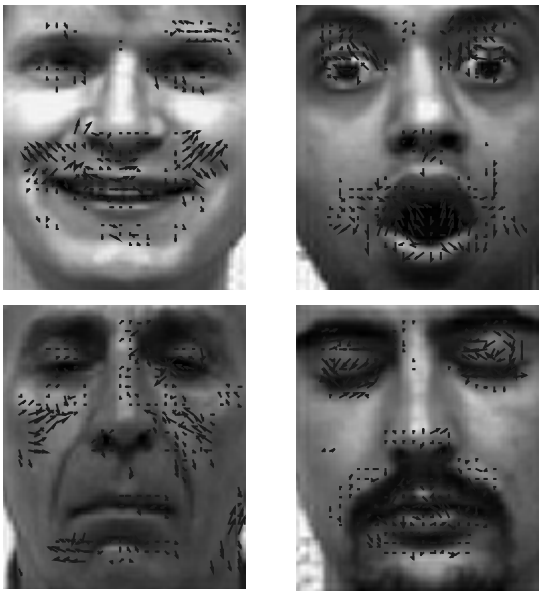

Figure 3. Estimated motion and the facial areas with substantial movement for the expressions "happy", "surprised" ", "sad" and "sleepy".

Figure 4 depicts the Radon transforms corresponding to expressions "happy", "surprised", "sad" and "sleepy", calculated at 0 and 90 degrees. Two different facial images are used for each expression, depicted with solid and dotted lines. It is evident that Radon transforms derived from images of two different persons look similar to each other if they correspond to the same expression. Conversely, transforms derived from two images of the same person with different expressions can be easily distinguished. Note also that since images of the same resolution have been used (31×28 image blocks) and image registration has been performed, Radon transforms are already aligned with each other, with the exception of a scale difference of expression "surprised" at 0^0 and a translation difference of "sleepy" at 90^0.

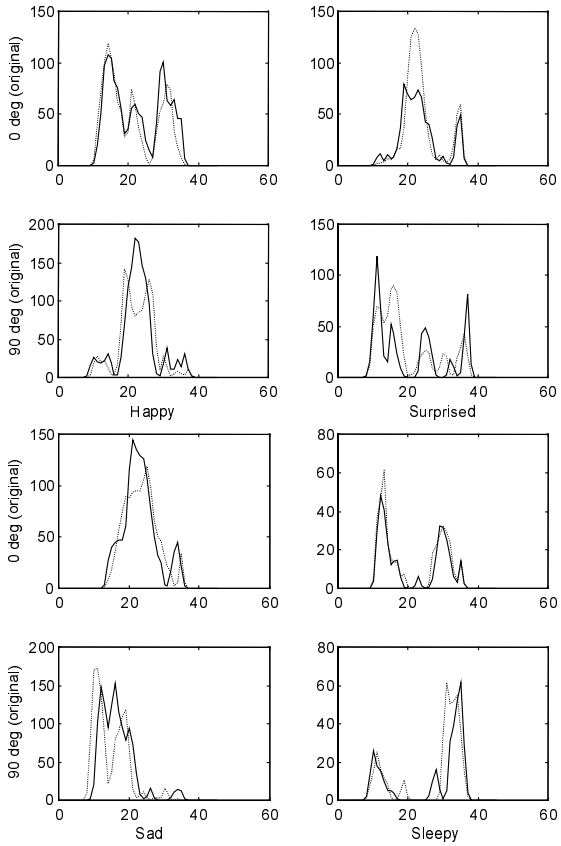

Figure 4. Radon transforms corresponding to expressions "happy", "surprised", "sad" and "sleepy", calculated at 0 and 90 degrees. Images of two different persons are used for each expression, depicted with solid and dotted lines.

The above misalignments are eliminated by using normalized signatures, illustrated in Figure 5 as vectors of length K=100. It is clear that all information regarding the shape of the Radon transform curves is retained, so that facial expressions can still be distinguished. Moreover, exact vertical scaling alignment is accomplished, whereas leading and trailing zeros are removed, resulting in exact horizontal scaling/translation alignment.

The importance of signature normalization is more evident from Figure 6, where images of two different persons at different resolutions are used. Radon transforms are represented by vectors of different lengths with scaling and translation differences in this case,

while normalized signatures can be directly used for comparisons. It should be noted that the two transform curves are not "matched" in any way, since normalization parameters of one curve are not dependent on any other curve; the same linear operations are applied to both.

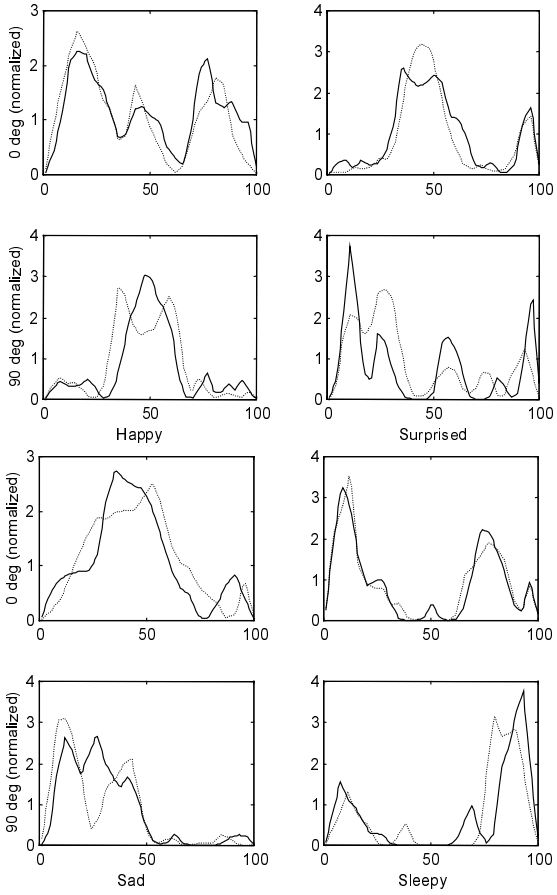

Figure 5. Normalized signatures corresponding to expressions "happy", "surprised", "sad" and "sleepy", again calculated at 0 and 90 degrees. The same images of two persons are used for each expression, depicted with solid and dotted lines.

Expression classification based on correlation coefficients between normalized signatures is first demonstrated in Table 1. Normalized signatures of length K=100 are obtained for angles 0^0 and 90^0 and concatenated into vectors of length 200. Correlation coefficients are then calculated between vectors derived from images of two different persons (A and B) at four different expressions "happy", "surprised", "sad" and "sleepy". It is clear that vectors corresponding to the same expression of different persons (shown as gray-shaded areas) produce high correlation coefficient values and can be distinguished, with the exception of "surprised B" and "sad A" which are misclassified as the same expressions.

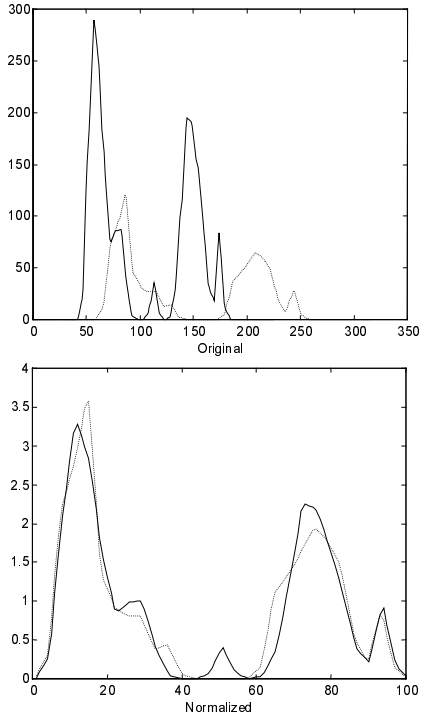

Figure 6. Original and normalized signatures corresponding to expression "sleepy" at 0 degrees. Images of two persons at different resolutions are used (solid and dotted lines).

Classification using Neural Networks

Although classification using simple correlation seems efficient we investigated also classification using a non-linear classifier. We modified a MLP network with four output units and one hidden layer to classify the normalized signatures of the four expressions. 20 signatures were used for the training set, 8 were used for the validation set and the remaining 32 were used for testing. After about 400 learning cycles, using a variant of the backpropagation algorithm, the network with 200 input units, 20 hidden units, and four output units converged, and was able to recognize all expressions from the training set and 7 out of 8 form the validation set. The network was also able to generalize well to the signatures of the test set (87.5% overall correct classification). The obtained results are summarized in the Table 2. Expressions "happy" and "sleepy" were perfectly classified while in contrast "sadness" obtained the poorest classification. The results are quite logical since "sadness" does not present a clear movement of some specific areas. In contrast the other three expressions involve movement of specific facial parts around mouth and eyes which distinguish them. Similar recognition rates were obtained using an LVQ network but we did not concentrate on it due to the small dataset which we were using.

Expression	Happy A	Happy B	Surprised A	Surprised B	Sad A	Sad B	Sleepy A	Sleepy B
Happy A	1.0000	0.7859	-0.1193	-0.1906	-0.0225	0.1323	-0.0194	0.0262
Happy B	0.7859	1.0000	-0.1619	-0.1461	0.0791	0.1750	-0.0550	-0.0107
Surprised A	-0.1193	-0.1619	1.0000	0.7564	0.6560	0.5561	-0.1663	-0.2771
Surprised B	-0.1906	-0.1461	0.7564	1.0000	0.8014	0.5475	-0.2752	-0.3388
Sad A	-0.0225	0.0791	0.6560	0.8014	1.0000	0.8411	-0.3254	-0.3196
Sad B	0.1323	0.1750	0.5561	0.5475	0.8411	1.0000	-0.2686	-0.2383
Sleepy A	-0.0194	-0.0550	-0.1663	-0.2752	-0.3254	-0.2686	1.0000	0.7842
Sleepy B	0.0262	-0.0107	-0.2771	-0.3388	-0.3196	-0.2383	0.7842	1.0000

Table 1. Correlation coefficients between normalized signatures derived from images of two persons (person A and person B) at expressions "happy", "surprised", "sad" and "sleepy".

Expression	Happy	Sad	Surprised	Sleepy
Happy	8	0	1	0
Sad	0	5	0	0
Surprise	0	2	7	0
Sleepy	0	1	0	8
Success	100%	62.5%	87.5%	100%

Table 2. Recognition rates obtained form the test set for the expressions "happy", "surprised", "sad" and "sleepy".

7. CONCLUSIONS - FURTHER WORK

An expression recognition scheme based on features derived from the optical flow between a neutral and a face at the 'apex' of the expression has been presented, utilizing Radon transform to estimate the basic movement within the face area. A technique for normalization of curves derived from the Radon transform in different angles is used, leading to simple classification either using correlation or a neural classifier.

Although the results are promising, the authors are currently applying the method on a larger database to evaluate its efficiency. In addition, an extension of the method is currently under investigation for expression recognition using video sequences. The necessity of estimating the 'apex' of the expression is thus eliminated and more temporal information is added, which is critical in many expressions.

8. ACKNOWLEDGMENTS

The present work is funded by the project PHYSTA (Principled Hybrid Systems: Theory and Applications, 1998-2001) of the Training Mobility and Research Program of the European Community. The authors are within the team of project PHYSTA, where speech and phychological cues are also used for emotion classification.

9. REFERENCES

[1]. R. Cowie and E. Douglas-Cowie, "Speakers and hearers are people: reflections on speech deterioration as a consequence of acquired deafness," In K-E. Spens and G. Plant (Eds) Profound deafness and speech communication, Whurr Publications, London, 510-527, 1995.

[2]. K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions*, vol. E74, 3474-3483, 1991.

[3]. Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(6), 636-642, 1996.

[4]. N. Intrator, D. Reisfeld and Y. Yeshrun, "Face Recognition using a Hybrid Supervised/Unsupervised Neural Network," *Pattern Recognition Letters* 17, 67-76, 1996.

[5]. N. Arad and D. Reisfeld, "Image Warping using few Anchor Points and Radial Functions," *Computer Graphics Forum*, vol. 14 (1), 35-46, 1994.

[6]. N. Tsapatsoulis, M. Leonidou and S. Kollias, "Facial Expression Recognition Using HMM with Observation Dependent Transition Matrix," *Proc. of MMSP'98,* Portofino, CA, December 1998.

[7]. Y. Xirouhakis, Y. Avrithis and S. Kollias, "Image Retrieval and Classification Using Affine Invariant B-Spline Representation and Neural Networks," *Proc. IEE Colloquium in Neural Nets and Multimedia*, London, Oct. 1998.

[8]. Z. Huang and F. S. Cohen, "Affine Invariant B-Spline Moments for Curve Matching," *IEEE Trans. Image Processing*, vol. 5 (10), Oct. 1996.