Mouth Center Detection under Active Near Infrared Illumination

THORSTEN GERNOTH, RALPH KRICKE, ROLF-RAINER GRIGAT
Hamburg University of Technology
Vision Systems, E-2
Harburger Schloßstr. 20, 21079 Hamburg
GERMANY

Abstract: Mouth center detection is important for speech and speaker recognition based on lip movements. In this paper we present an algorithm for mouth center detection under active near infrared illumination. Face images captured in the near infrared spectrum are normalized with respect to size and in-plane rotation of the eyes. A coarse image region containing the mouth can be determined in the normalized images. We present an approach for locating the center of the mouth using morphological operations and amplitude projection of the mouth region. Experimental results demonstrate the effectiveness of the presented mouth center detection algorithm under simulated realistic conditions.

Key–Words: Infrared imaging, Visual speech recognition, Mouth detection

1 Introduction

Today, reliable and automatic person identification is needed for many domains. Using biometric data to identify a target person has some well known conceptual advantages. For example, the identification procedure is immutable bound to the person which should be identified. For the access control to security areas, a person identification system should be hard to deceive by possible imposter attacks.

Using face images as a biometric characteristic has gained much attention [1, 2]. Most face recognition systems identify a person based on still face imagery. This does not provide for an adequate deterrent against imposter attacks, such as presenting a prerecorded still image to the system. To resist such attacks, the liveness of the target person can be verified by searching for movement in successive images, such as repetitive eyelid or lip movement. However, to account only for movement does not guard against attacks by use of a prerecorded video or a mask of a still image with moving lips. Luettin et al. [3] showed that the movements of the lips contain speech dependent information as well as speaker dependent information. By using speaker dependent information as an additional biometric feature to identify a person, imposter attacks such as a mask with moving lips can be ruled out. Considering also the speech dependent information, e. g. reading a pass-phrase from the lips of the target person, can make it difficult to deceive the system with prerecorded video.

A system for speech and speaker recognition based on lip movements needs to solve several tasks (Fig. 1). In general, detection of the mouth region and tracking in successive images are required. Detection and tracking of the mouth is a difficult problem due to the large number of varying parameters and conditions. The mouth shape is not only person dependent but also constantly deformed, e. g. due to speech production, breathing, and distortions such as head pose rotations.

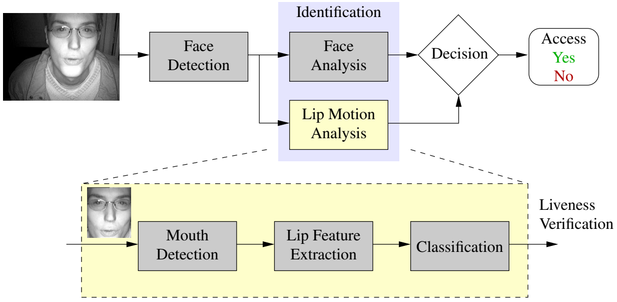


Figure 1: Lip motion analysis for visual speech and speaker recognition

The system should rely on video in the infrared spectrum as the only sensory modality of the system. Infrared spectrum imagery has the well known advantage of being insensitive to changes of the visible illumination. A requirement for the mouth detection is that it should work person independently. Appearance based features for visual speech and speaker recognition can be extracted from a mouth region-of-interest [4]. An appropriate region-of-interest can be determined from the mouth center.

Various algorithms to detect the mouth or lips in images and to track them in successive images have been proposed. For example in [5] an approach to locate and track lips in gray-scale image sequences is described. A deformable model of the lips is used to guide the search for lips in the images. The model is gradually moved and deformed within the image to find the best match. Shape and appearance can also be combined in the model to locate and track the lips, e. g. in Active Appearance Models [6]. Another approach is to use templates to find facial features in the mouth region. In [7] templates are used to find the left and right mouth corner point in infrared images. The best results are reported with person dependent templates. The performance significantly degrades when person independent templates are used. If color information is available, detecting the lips in images is simplified. Lips are predominantly red. This can be used to segment the lip region from face images. In [8] the image is hue-filtered and thresholded afterwards such that a binary image is obtained. The mouth position can then be estimated by analyzing the spatial distribution and the dispersion along the horizontal and vertical axis of the pixels above the threshold. An estimate of the outer mouth corner points is obtained by projecting the vertical gradient on the horizontal and vertical axis. The horizontal and vertical projection can also be used to locate the mouth corner points in gray-scale images [9, 10].

In the next section, an overview of the system is given. In Section 2.1 the database, which is used for all experiments, is briefly described. The preprocessing procedure, used to normalize the images, is outlined in Section 2.2. The proposed algorithm to determine the mouth center will be presented in Section 3. Experimental results are shown in Section 4. Finally, the last section concludes this work.

2 System Overview

2.1 Database

We use the TUNIR [11] database for all experiments. The database consists of recordings of 74 people speaking English digits in front of a camera. The subjects were asked to look into the camera while they speak English digits from zero to nine. There are four sequences of every subject. In two of the sequences, the subjects wear glasses. Most of the subjects in the database are between 20 and 30 years of age. There are both male and female subjects. The ethnic background of the subjects is diverse; such as there are white skin, dark skin and Asian looking subjects in the database. Some subjects wear a beard.

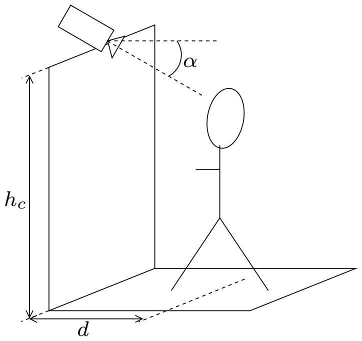


Figure 2: Camera setup

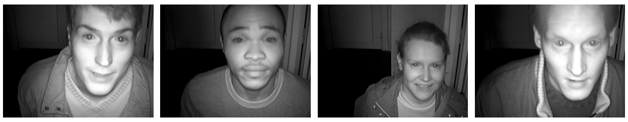


Figure 3: Example images from the TUNIR database

The subjects are recorded under near infrared illumination ($780-1100\,\mathrm{nm}$). The active infrared source is mounted on axis close to the camera. They stand at a distance between $0.39\ \mathrm{m} \leq d \leq 0.78\ \mathrm{m}$ (Fig. 2). Because of different distances of the subjects to the camera and anatomical differences, the head sizes in the recorded images vary. There are also illumination variations, mainly because of different distances of the subjects to the camera and therefore to the active infrared source.

Videos are recorded with a frame rate of 30 fps, a resolution of 320 × 240 and 8 bit per pixel. Some example images are shown in Fig. 3.

2.2 Normalization of Face Images

The images from the camera are normalized with respect to size and in-plane rotation of the eyes such that the eyes in the normalized images are at fixed positions and the distances between the eyes are constant. Localization of the eyes is therefore required. The eye localization makes use of the bright pupil effect under near infrared illumination [12]. The localized eye position corresponds to the pupils of the subjects in the images.

For visual speech and speaker recognition, the mouth area needs to cover a sufficient large region in the images to be able to extract features. The normalization with respect to the size of the images is based on the geometry of faces.

The face geometry was determined from 160 images of 14 subjects. The mouth corner and eye positions were manually labeled. Some insight about the geometry of the face images was gained from the labeled positions. Of interest was the distance between

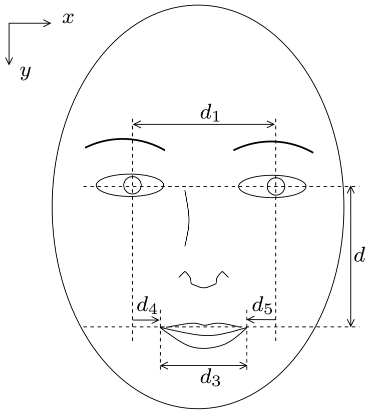


Figure 4: Distances between different facial features

	d_{max_i}	d_{min_i}	$\overline{d_i}$	σ_i
d_1	82.05	38.00	57.73	12.71
d_2	94.54	38.08	64.49	15.69
d_3	59.07	23.08	42.30	7.58

Table 1: Distances between facial features [pixel]; original images

the eyes d_1, the distance between the mouth corner points d_3 and the distance between the center of the eyes and the center of the mouth d_2 (Fig. 4). The determined distances can be found in Table 1. Based on these measurements, the conclusion was drawn to normalize the distance between the eyes to the mean distance $\overline{d_1} = 58$ pixels. All further experiments are based on this normalization distance. As written above, the eyes in the normalized images are at a fixed position. Some statistical properties of the normalized faces can be found in Table 2. Additionally, the horizontal distance between the left eye and the left mouth corner d_4 and respectively the right eye and the right mouth corner d_5 is given (Fig. 4).

3 Mouth Center Detection

Based on the examination of the face geometry of the recorded subjects, some heuristics to detect the mouth center can be developed. Based on the evaluated distances in Table 2, a coarse mouth region of size 58×64 pixels can be determined. As described in Section 2.2, the eyes in the normalized images are at the same vertical position. The horizontal distance between left and right eye is fixed to $d_1 = 58$ pixel. From Table 2 one can observe, that the mouth is displaced from the horizontal center of the eyes by up to

	d_{max_i}	d_{min_i}	$\overline{d_i}$	σ_i
d_2	76.00	50.50	63.86	4.48
d_3	58.07	28.01	42.85	5.75
d_4	15.00	-1.00	6.97	3.03
d_5	20.00	-3.00	8.20	3.66

Table 2: Distances between facial features [pixel]; normalized eye distance $d_1 = 58$ pixel

$$d_h = -\min\{d_{min_4}, d_{min_5}\} = 3 \text{ pixels.}$$

In vertical direction, the dispersion of the mouth corners from the mean $\overline{d_2} = 64$ pixel is up to

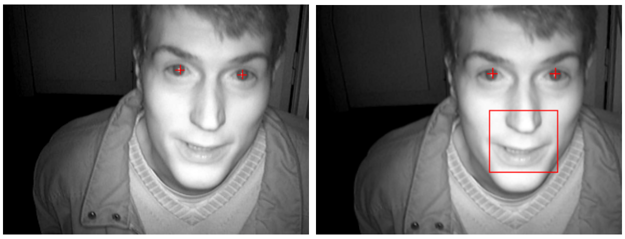
$$d_v = \max\{\overline{d_2} - d_{min_2}, d_{max_2} - \overline{d_2}\} = 13 \text{ pixels.}$$

Assuming that the expansion of the mouth from the center in vertical direction is not larger than $d_e = 16$ pixels, we extract a window with the pixel coordinates x_{roi} and y_{roi} as follows (x_{eye_l} is the horizontal coordinate of the left eye and y_{eye_l} the vertical coordinate, respectively):

$$\begin{array}{rrl} & x_{roi} \geq & x_{eye_l} - d_h \\ \wedge & x_{roi} \leq & (x_{eye_l} + d_1) + d_h \ , \\ & y_{roi} \geq & (y_{eye_l} + \overline{d_2}) - (d_v + d_e) \\ \wedge & y_{roi} \leq & (y_{eye_l} + \overline{d_2}) + (d_v + d_e) \end{array}$$

The extracted region is a rectangular part of the image containing the lips, surrounding skin including parts of the nose, and even some non-facial background, as can be seen in Fig. 5. The non-facial background is generally present in the extracted mouth regions since the subjects are recorded slightly from above and the camera is not centered to the lips of the subjects. The background appears as a dark structure at the lower corners of the extracted rectangular region. The background is masked by an adaptive threshold applied to the extracted region. To account for the differences of illumination the extracted region is additionally normalized through histogram equalization.

The horizontal center of the mouth is determined by the assumption that it is approximately on the normal through the center of the line connecting the two eyes. The vertical center is determined by a horizontal projection of the gray level amplitudes of the mouth region [13]. Amplitude projection is a technique where all pixel values in a row or column of a rectangular image are summed. Amplitude projection proved to be useful for facial feature extraction in the visible domain [14]. Here we apply this technique to



(a) Original image with tracked eye positions (b) Normalized image with normalized eye positions and mouth region-of-interest

Figure 5: Extraction of coarse mouth region-of-interest based on geometric considerations

near infrared face images. The amplitude projection of image $\boldsymbol{I}(x,y)$ onto its vertical axis (horizontal projection) is defined as (x and y are pixel coordinates and N_x the number of horizontal pixels):

$$\boldsymbol{H}(y) = \sum_{x=1}^{N_x} \boldsymbol{I}(x,y)$$

Amplitude projection onto the vertical axis can be used to determine dominant horizontal structures, e. g. vertical gradients in the images. The mouth constitutes the darkest horizontal structure in the near infrared images. Generally the teeth are occluded by the upper lip or there is shadow on the teeth. The teeth do not appear very bright in the recorded images.

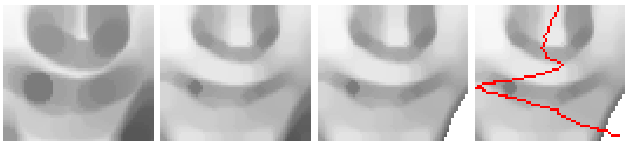
To enhance the detection of the vertical mouth center, the morphological erosion followed by dilation is applied to the images prior to the horizontal projection. With erosion small objects in the image can be removed. The gray-scale erosion of image $\boldsymbol{I}$ with structuring element $\boldsymbol{B}$ is defined with $(x'+x, y'+y) \in$ domain of I and $(x,y) \in$ domain of B as:

$$(\boldsymbol{I} \ominus \boldsymbol{B})(x',y') = \min\left\{\boldsymbol{I}(x'+x, y'+y) - \boldsymbol{B}(x,y)\right\}$$

The erosion operation removes bright details from the images. The size of the structuring element of erosion is 5×5. Subsequent to the erosion, dilation is applied to the images. Gray-scale dilation is defined with $(x'-x, y'-y) \in$ domain of I and $(x,y) \in$ domain of B as:

$$(\boldsymbol{I} \oplus \boldsymbol{B})(x',y') = \max\left\{\boldsymbol{I}(x'-x, y'-y) + \boldsymbol{B}(x,y)\right\}$$

A structuring element of size 3×3 is used with the dilation operator. The processing of the images with the morphological operators smoothes the lip contours and removes small bright details from the images. The mouth appears as connected and dominant horizontal structure in the images. The vertical mouth center can then be estimated by the horizontal projection of the mouth region. The successive processing steps to locate the mouth center are summarized in Fig. 7.



(a) Erosion (b) Dilation (c) Masking of background (d) Horizontal projection

Figure 6: Successive image processing to estimate the vertical mouth center. Horizontal projection $\boldsymbol{H}(y)$ is plotted as red line on the horizontal axis.

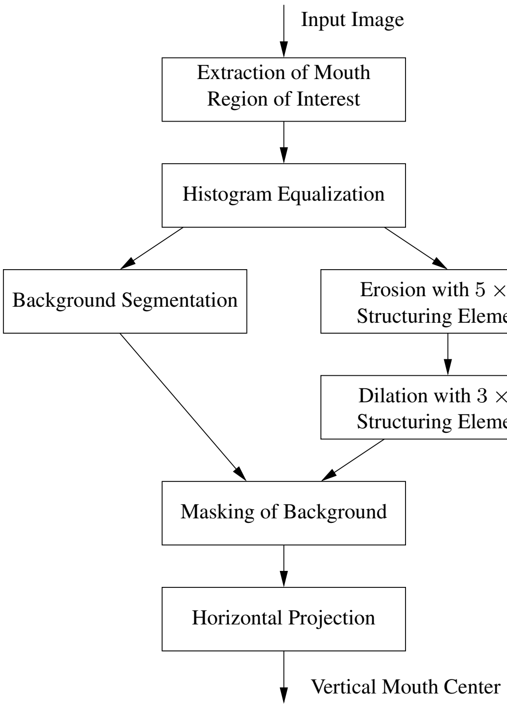


Figure 7: Estimation of vertical mouth center

4 Experiments

To evaluate the performance of the mouth center detection algorithm described in Section 3, estimated mouth centers are compared against manually labeled positions. In images of 14 different subject of the TU-NIR database, the mouth corner points were manually labeled. The mouth center is assumed to be the center point of the manually labeled mouth corners. The *true* mouth center is determined from the mouth corner points since the mouth corners constitute more prominent features in the images. Manually labeling

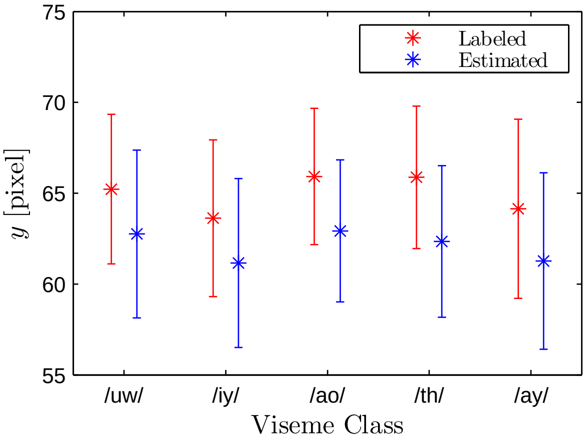


Figure 8: Labeled and estimated vertical position of mouth center with respect to different viseme classes. The mean is marked with a cross, the length of each bar shows the standard deviation.

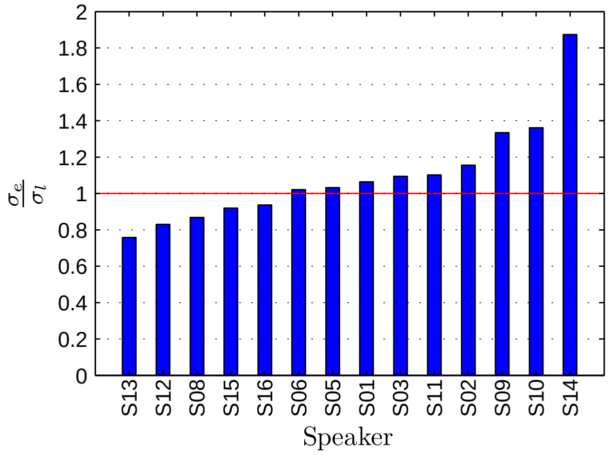


Figure 9: Standard deviation of estimated vertical mouth positions σ_e related to standard deviation of respective labeled positions σ_l for different subjects

the mouth center seemed to be too difficult. However, manually labeling of the mouth corners is also prone to errors due to the small size of the images and due to smooth transitions between the inner and outer lip contour.

The algorithm is evaluated with respect to different speakers and with respect to different mouth shapes (visemes). As stated above, the subjects speak the digits from zero to nine in every sequence of the TUNIR database. From these sequences, static images corresponding to 5 different visemes were extracted to test the performance of the algorithm. The 5 different viseme classes are:

- /uw/, articulation with lip rounding, narrow and round lips, extracted form the word *two*
- /iy/, articulation with lip rounding, narrow lips, extracted form the word *three*
- /ao/, articulation with lip rounding, wide open and round lips, extracted form the word *four*
- /th/, dental articulation, extracted form the word *three*
- /ay/, articulation with lip rounding, open lips, extracted form the word *nine*

In Fig. 8, the estimated vertical pixel coordinates of the mouth center and the respective pixel coordinates of the labeled positions are plotted for different viseme classes. The coordinates are relative to the left eye to make them independent of position changes of the subject in the images.

There is a slight offset on the estimated mouth center positions for the different viseme classes. This difference of the mean of the coordinates of the labeled and estimated positions is about three pixels. This offset results from calculating the *true* mouth center from the middle of the labeled mouth corner positions. This may not be optimal for all lip shapes. The smallest offset is for viseme classes /iy/ and /uw/. The shape of the lips of these viseme classes is usually narrow and round. The mouth corners are rather well aligned with the mouth center.

Generally it is difficult to conclude on the performance of the algorithm from a sequence of images, but it can be expected, that for the same sequence the variability of estimated mouth positions is equivalent to the variability of manually labeled mouth positions. We relate the standard deviation of the estimated mouth positions σ_e to the standard deviation of the labeled mouth positions σ_l. In Fig. 9, the ratio $\frac{\sigma_e}{\sigma_l}$ is plotted with respect to different speakers. One can see that the standard deviations differ significantly for subject $S09$, $S10$ and $S14$ of the TUNIR database. The variability of the estimated mouth positions of subject $S09$ and $S10$ is of the same magnitude than for other subjects. Noticeable is that the lips of subjects $S09$ and $S10$ are very well contoured. We assume that manually labeling the mouth corners is therefore more accurate. Subject $S14$ of the TUNIR database wears a very prominent dark beard. This is a problem for the algorithm. As it can be seen in Fig. 10, there is not always a unique minimum in the horizontal projection of the pixels of the mouth region.

5 Conclusion

An algorithm to detect the mouth center under active near infrared illumination was presented. The perfor-

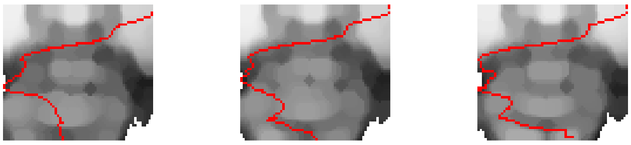


Figure 10: Horizontal projection $\boldsymbol{H}(y)$ of mouth region of subject wearing dark beard. $\boldsymbol{H}(y)$ is plotted as red line on the horizontal axis.

mance of the algorithm was compared against manually labeled features in face images. The accuracy of the detected mouth center positions is generally within the variability of manually labeled features in the images. Problems are caused by subjects with prominent dark beards. The proposed algorithm proved to be useful to extract a region-of-interest of the mouth area from face images and is currently used in a visual speech and speaker recognition system.

Acknowledgement

This work is part of the project *KATO – Innovative Cabin Technologies* funded by the German Federal Ministry of Economics and Technology.

References:

[1] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, Face Recognition: A Literature Survey, *ACM Computing Surveys*, Vol. 35, No. 4, 2003, pp. 399–458.

[2] S. Zhao and R.–R. Grigat, An Automatic Face Recognition System in the Near Infrared Spectrum, In *Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2005)*, 2005, pp. 437–444.

[3] J. Luettin, N. A. Thacker and S. W. Beet, Speaker Identification By Lipreading, In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, 1996, pp. 62–65.

[4] G. Potamianos, C. Neti, J. Luettin and I. Matthews, Audio-Visual Automatic Speech Recognition: An Overview, In E. Vatikiotis–Bateson, G. Bailly and P. Perrier (eds.), *Audio-Visual Speech Processing*, MIT Press, 2004.

[5] J. Luettin and N. A. Thacker, Speechreading using Probabilistic Models, *Computer Vision and Image Understanding*, Vol. 65, No. 2, 1997, pp. 163–178.

[6] I. Matthews, T. F. Cootes, J. A. Bangham, S. C. and R. Harvey, Extraction of Visual Features for Lipreading, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, 2002, pp. 198–213.

[7] F. Shafait, R. Kricke, I. Shdaifat and R.–R. Grigat, Real Time Lip Motion Analysis for a Person Authentication System Using Near Infrared Illumination, In *Proceedings of the 2006 IEEE International Conference on Image Processing (ICIP 2006)*, 2006, pp. 1957–1960.

[8] T. Coianiz, L. Torresani and B. Caprile, 2D Deformable Models for Visual Speech Analysis, In D. G. Stork and M. E. Hennecke (eds.), *Speechreading by Humans and Machines: Models, Systems and Applications*, Springer-Verlag, Berlin, 1996, pp. 391–398.

[9] R. Stiefelhagen, U. Meier and J. Yang, Real-Time Lip-Tracking for Lipreading, In *Proceedings of the Eurospeech '97*, 1997, pp. 2007–2010.

[10] D. Shah and S. Marshall, Image Models for Facial Feature Tracking, In C. Kotropoulos and I. Pitas (eds.), *Nonlinear Model-Based Image/Video Processing and Analysis*, John Wiley & Sons, Inc., 2001, pp. 299–319.

[11] S. Zhao, R. Kricke and R.–R. Grigat, TUNIR: A Multi-Modal Database for Person Authentication under Near Infrared Illumination, To be presented at *6th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA 2007)*, 2007.

[12] S. Zhao and R.–R. Grigat, Robust Eye Detection under Active Infrared Illumination, In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, 2006, pp. 481–484.

[13] W. K. Pratt, *Digital Image Processing (3rd ed.)*, John Wiley & Sons, Inc., 2001.

[14] M.–H. Yang and D. J. Kriegman and N. Ahuja, Detecting Faces in Images: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, 2002, pp. 34–58.