OBJECT TRACKING BASED ON MULTISCALE MORPHOLOGICAL TEMPLATES

G.N. Stamou, N. Nikolaidis and I. Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, 54124 Thessaloniki, GREECE
phone: + (30) 2310996361, fax: + (30) 2310996304, email: {gstamou,nikolaid,pitas}@aiia.csd.auth.gr
web: http://www.aiia.csd.auth.gr

ABSTRACT

This paper presents a novel template representation that can be used in template-based object tracking methods. More specifically, the Multiscale Morphological Template is introduced and incorporated in a template-based object tracking algorithm. The proposed template can be updated over time to cope with changing environment/object conditions. The algorithm is applied to face tracking in scenes with complex background. Results of the object tracking algorithm using the proposed and existing template representations are compared using measures based on ground truth data. The proposed template is proved to be superior to existing templates.

1. INTRODUCTION

Video-based object tracking has received considerable attention by the research community in the past decades, mainly due to the wide range of its potential applications. A great amount of this research has focused on tracking humans for applications such as smart surveillance, human computer interaction, motion capture for animation, coding, compression, content-based querying, indexing and retrieval, gesture recognition, and 3D reconstruction. However, the difficulties introduced by different capturing, object and environment conditions, such as the use of a single camera and the associated projection ambiguities, poor or varying lighting conditions, occlusion or self occlusion, unconstrained motion, clutter, variability in the body shape and appearance of humans due to their clothing etc., has forced researchers to make a number of assumptions in order to handle specific aspects of the overall problem. These can be related to the camera/object motion, to the environment conditions or the object conditions. For a comprehensive review of different methods, the reader is referred to [1] and [2].

Template matching techniques have been used by many researchers to perform 2D object tracking. The first step (initialization step) in template-based algorithms involves the selection of the template that will be used, i.e. the creation of an image model of the object to be tracked. Such models can be individualized and acquired on-line (e.g. using the first frame of the video sequence) or they can be generic and created off-line by employing statistical methods. Face templates can be obtained e.g. by the use of eigenfaces or Gabor wavelets, both of which have been used in many tasks, including object/face recognition, verification, authentication and tracking [3, 4]. Object tracking based on template matching involves searching the current frame of the video sequence to determine the image region that best resembles the template, based on a similarity measure. The Sum of Absolute Differences (SAD) and the Sum of Squared Differences (SSD) are frequently used as similarity measures because of their simplicity, while others, such as joint entropy, normalized correlation and mutual information have also been employed [1].

In this paper, a novel template representation for use in 2-D template-based object tracking methods is presented. The object of interest is represented by a multivalue template, the Multiscale Morphological Template. Vectors are assigned to each template pixel, by applying the multiscale morphological dilation-erosion using a scaled structuring function [5] to the image points that correspond to the template pixels. The motivation that led us to seek an alternative to the Gabor filters tuned to different orientations and scales, that have been previously used to build the vectors in the templates, was mainly their increased computational overhead. Additionally, multiscale morphological techniques provide good object representations (e.g. representation of facial features and faces), since dilations-erosions deal with local extrema in the image. This has been verified in other applications, such as face recognition and authentication [6], where the Gabor analysis has been superseded by the multiscale morphological dilation-erosion. The template can be initialized as the bounding or the escribed box (to avoid the inclusion of background pixels in the template) of the object of interest (e.g. a human face) in the first frame of a video sequence. Initialization can be manual, as in this paper, or automatic (i.e. the output of an object detection module).

The location of the tracked object in the current frame is calculated based on the maximization of the similarity of the vectors at the corresponding pixels in the template and the image pixels on all possible template locations in the current frame of the video sequence. If the similarity of the template falls below a threshold, the initial template used in the matching process is replaced by the image region that corresponds to the template location in the previous frame.

The object tracking algorithm used to test the proposed template representation can be classified as a standard template-based rigid object tracking method. Assumptions made include presence of a single object in the scenes, partial occlusion or no occlusion at all and known initial position of the object of interest. The algorithm can track objects in scenes with uncontrolled lighting conditions and a complex/moving background. The results of the object tracking algorithm using the proposed template are compared against the results obtained with existing template representations. The comparison is performed using ground truth-based performance measures and the proposed template representation is proved to be superior to existing template representations.

The remainder of the paper is organized as follows. Section 2 describes the Multiscale Morphological Template representation and incorporates it into a standard template-based object tracking algorithm. In Section 3, variants of existing template representations and object tracking algorithms that are compared with the proposed method are briefly described. Section 4 presents the ground truth data-based measures that are used to evaluate the performance of the object tracking algorithm when using the proposed and existing template representations. In Section 5, experimental results are presented. Conclusions are drawn in Section 6.

2. MULTISCALE MORPHOLOGICAL TEMPLATE TRACKING

As already mentioned, the goal of tracking based on template matching is to search the current frame of the video sequence in order to determine the image region that best resembles the template, based on a similarity measure. Essentially, the goal of a template matching algorithm is to estimate the parameters of a geometrical coordinate transformation, which would "place" the template onto the target image in such a way as to maximize (minimize) the similarity (distance) measure used. Let the superscripts t and c de-

note the template and the current frame respectively and let A^c be the image region corresponding to the object being tracked in the current frame c. For a rigid object, A^c can be obtained from the template, denoted by A^t, by employing a coordinate transformation $\theta(A^t) \Rightarrow A^c$ the parameters of which should be estimated by the algorithm. Affine (rotation, translation, scaling) or quadratic transformations can be employed. Therefore, every point $\mathbf{x}^c(x,y)$ in the target region is obtained from a corresponding point $\mathbf{x}^t(x,y)$ in the template:

$$\mathbf{x^c} = \theta(\mathbf{x}^t; \mathbf{a}^c), \tag{1}$$

where $\mathbf{a}^c$ denotes the transformation parameter vector associated with A^c and determines the object location in the current frame. Estimation of the transformation parameters is performed by identifying the image region that best matches the template.

Let R and Z denote the set of real and integer numbers, respectively. Given an image $f(\mathbf{x}) : F \subseteq Z^2 \rightarrow R$ and a structuring function $g(\mathbf{x}) : G \subseteq Z^2 \rightarrow R$, the grayscale dilation and erosion of the image $f(\mathbf{x})$ by $g(\mathbf{x})$ are denoted by $(f \oplus g)(\mathbf{x})$ and $(f \ominus g)(\mathbf{x})$ respectively. The multiscale dilation-erosion of the image $f(\mathbf{x})$ by $g(\mathbf{x})$ is defined by [5]:

$$(f \star g_\sigma)(x) = \begin{cases} (f \oplus g_\sigma)(\mathbf{x}) & \text{if} \quad \sigma > 0 \\ f(\mathbf{x}) & \text{if} \quad \sigma = 0 \\ (f \ominus g_{|\sigma|})(\mathbf{x}) & \text{if} \quad \sigma < 0 \end{cases} \tag{2}$$

where the integer σ denotes the scale parameter of the structuring function. The computational complexity of the dilation-erosion operations depends on the choice of the structuring function. A circular structuring function is often employed for fast computation of these morphological operations, as in this paper. Figure 1 depicts the output of multiscale dilation-erosion when applied to a facial template for various values of the scale parameter. It can be easily seen that the multiscale dilation-erosion captures significant information with respect to distinctive features such as the eyes, eyebrows, nose tip, nostrils, lips, face contour, etc.

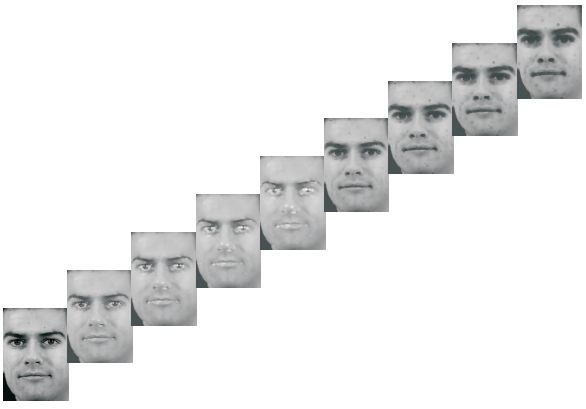

Figure 1: The multiscale dilation (sub-images 2-5) and erosion (sub-images 6-9) of a facial image (sub-image 1), using a circular structuring function for scale parameter values 1-4 respectively.

Let $V^t = \{\mathbf{x}_{k,l}\}$, $k = 1,...,M$, $l = 1,...,N$ be the set of pixels of an $M \times N$ rectangular template. The outputs of multiscale dilation-erosion for $\sigma = -\sigma_m,...,\sigma_m$ form the vector $\mathbf{j}$, also called "jet":

$$\begin{aligned} \mathbf{j}(\mathbf{x}_{k,l}) = ((f \star g_{\sigma_m})(\mathbf{x}_{k,l}),...,(f \star g_1)(\mathbf{x}_{k,l}), f(\mathbf{x}_{k,l}), \\ (f \star g_{-1})(\mathbf{x}_{k,l}),...,(f \star g_{-\sigma_m})(\mathbf{x}_{k,l})) \end{aligned} \tag{3}$$

where $\mathbf{x}_{k,l} = (k,l)$ denotes the coordinates of a template pixel.

The Multiscale Morphological Template Tracking (MMTT) algorithm tracks an object by exhaustively searching the current frame in a region around the location of the template in the previous frame, for the template location, that maximizes (minimizes) a similarity (distance) function between the vectors at the template pixels and the corresponding image pixels of the candidate template locations. The exhaustive search is bounded by a maximum offset in the x and y directions, which is set to 30 pixels in this paper. To cope with scale changes of the tracked object (e.g. in case a person is moving towards/away from the camera), the exhaustive search in the current frame is repeated for different template sizes, which are determined by the size of the tracked object in the previous frame. To cope with in-plane rotation, the template should be rotated as well, i.e. template rotations are not considered. In the current implementation, the algorithm can cope with object translation parallel to the camera and towards/away from the camera. The similarity function is based on the norm of the difference between the vectors that correspond to each template pixel and the corresponding image pixel of the candidate template locations in the current frame. Let $V^c = \{\mathbf{x}^c_{r,s}\}$ be the set of image pixels of a candidate ($M \times N$) template location in the current frame, where $\mathbf{x}^c_{r,s} = (r,s)$ denotes the coordinates of an image pixel; $r = r_0 + k$ and $s = s_0 + l$, where (r_0, s_0) denotes the image coordinates of the top-left pixel of a candidate template location, $k = 1,...,M$ and $l = 1,...,N$. The goal of the MMTT algorithm is to find the template location in the current frame, where the set V^c of its image pixel coordinates maximizes (minimizes) the similarity (distance) function:

$$S(\{\mathbf{x}^c_{r,s}\}) = \sum_{(k,l) \in V^t} \left\{ S_v(\mathbf{j}(\mathbf{x}^c_{r,s}), \mathbf{j}(\mathbf{x}^t_{k,l})) \right\} \tag{4}$$

where $S_v(\mathbf{j}(\mathbf{x}^c_{r,s}), \mathbf{j}(\mathbf{x}^t_{k,l}))$ denotes a similarity (distance) measure between the vectors at template pixel (k,l) and the corresponding image pixel (r,s) of the candidate template locations in the current frame. Two measures have been employed in the similarity (distance) function of the MMTT algorithm, namely the normalized correlation:

$$S_v(\mathbf{j}(\mathbf{x}^c_{r,s}), \mathbf{j}(\mathbf{x}^t_{k,l})) \triangleq \frac{\mathbf{j}(\mathbf{x}^c_{r,s}) \cdot \mathbf{j}(\mathbf{x}^t_{k,l})}{\|\mathbf{j}(\mathbf{x}^c_{r,s})\| \, \|\mathbf{j}(\mathbf{x}^t_{k,l})\|} \tag{5}$$

where $(\cdot)$ denotes the inner product, and the L_2 norm:

$$S_v(\mathbf{j}(\mathbf{x}^c_{r,s}), \mathbf{j}(\mathbf{x}^t_{k,l})) \triangleq \|\mathbf{j}(\mathbf{x}^c_{r,s}) - \mathbf{j}(\mathbf{x}^t_{k,l})\|. \tag{6}$$

To account for changes in the environment conditions (e.g. illumination changes) and changes caused by the object motion (e.g. the in-depth rotation of a human face), the template A^t is updated if the similarity (distance) between the template and the image region that matches the template in the current frame falls below (rises above) a threshold. In such a case, the initial template used in the matching process is replaced by the image region that best matched the template in the previous frame, which is employed in subsequent matching. Obviously, higher (lower) values of the similarity (distance) threshold should account for even small changes in the object/environment conditions. In the case of normalized correlation, the threshold is fixed, whereas in the case of the L_2 norm, the threshold varies. The normalized correlation and the L_2 norm threshold values that were used in the experiments presented in this paper were set to 0.5 and 75% of the mean value of the L_2 norm that produced the best template match in all previous frames respectively.

3. VARIANTS OF EXISTING TEMPLATE REPRESENTATIONS AND TRACKING ALGORITHMS

In order to facilitate comparison with other template representations, the algorithm described in Section 2 is modified so that 2-D Gabor wavelets (planar sinusoids multiplied by a two dimensional Gaussian) are used instead of (3) to form vectors at each template pixel. The vectors consist of coefficients that are computed by convolving the template pixel (k,l) with the following wavelet kernel:

$$W(k,l,\lambda,\theta,\phi,\mu,\gamma),= e^{-\frac{k'^2+\gamma^2 l'^2}{2\mu^2}}\cos(2\pi\frac{k'}{\lambda}+\phi) \quad (7)$$

where

$$k' = k\cos\theta + l\sin\theta, \qquad l' = -k\sin\theta + l\cos\theta. \quad (8)$$

λ specifies the frequency of the wavelet, θ specifies the orientation of the wavelet, ϕ specifies the phase of the sinusoid, μ specifies the radius of the Gaussian and γ is the aspect ratio of the Gaussian. Three different sets of parameter values have been employed, the original and a simplified set of the values introduced in [7] and the values proposed in [8]. These variants will be referred to as Gabor Template Tracking (GBTT) algorithms.

Two other simple template matching object tracking algorithms employing simple templates of grayscale values were also compared with the proposed method. The template matching step consists of exhaustively searching (similarly to the algorithm in Section 2) for the template location in the current frame of the video sequence that minimizes the SSD:

$$SSD = \sum_{(k,l)\in V^t} (A^c_{r,s} - A^t_{k,l})^2, \quad (9)$$

or maximizes the normalized correlation:

$$NC = \frac{\sum_{(k,l)\in V^t}(A^t_{k,l}-\bar{A}^t)(A^c_{r,s}-\bar{A}^c)}{\sqrt{\sum_{(k,l)\in V^t}(A^t_{k,l}-\bar{A}^t)^2\sum_{(k,l)\in V^t}(A^c_{r,s}-\bar{A}^c)^2}}, \quad (10)$$

between the template and the image region that corresponds to candidate template locations. $A^t_{k,l}$ and $A^c_{r,s}$ are the brightness values of the template pixel (k,l) and the corresponding image pixel (r,s), whereas $\bar{A}^t$ and $\bar{A}^c$ are the mean brightness values in the template and the image regions that correspond to candidate template locations in the current frame respectively. These two algorithms will be referred to as Grayscale Template Tracking (GTT)-SSD and (GTT)-NC from this point onwards. In all variants, the location of the object of interest was manually initialized and the template was updated in a manner similar to the algorithm presented in Section 2.

4. EVALUATION OF TRACKING PERFORMANCE

Performance evaluation of the proposed tracking algorithm was accomplished using manually acquired ground truth data. For this reason, the bounding box of human faces was manually outlined in all the frames of several video sequences. Two different error measures were employed to perform the evaluation. The first error measure calculates the percentage of the ground truth object image region that is not correctly tracked by the algorithm. The second error measure evaluates the percentage of the image region tracked by the algorithm that does not correspond to the actual object as defined by the ground truth data (i.e. the one that corresponds to portions of the background). Let S_g and S_a denote the ground truth image region and the image region tracked by the algorithm in a frame respectively. Then, the two measures can be defined on a frame basis by:

$$d_1 = \frac{A(S_g \bigcap S^c_a)}{A(S_g)}, \; d_2 = \frac{A(S_a \bigcap S^c_g)}{A(S_a)}, \quad (11)$$

and on a sequence-basis by:

$$D_i = \frac{1}{K}\sum_{t=1}^{K} d_i(t), \qquad i = 1,2 \quad (12)$$

where S^c_a denotes a set's complement, A denotes the area of an image region and K is the number of frames in the video sequence. These measures can be subsequently combined using a weighting scheme to produce a single numerical measure:

$$D = \alpha D_1 + \beta D_2 \quad (13)$$

where the parameters α and β are weight constants. The results presented in this paper were produced using $\alpha = \beta = 0.5$.

5. EXPERIMENTAL RESULTS

A series of tests were first conducted to compare the two similarity (distance) measures. Normalized correlation was proved to produce the best results. Another set of tests were then performed to assess the optimum length of the vectors (jets) formed at each template or image pixel, both in terms of accuracy and speed. Table 1 presents the results with respect to the error measures introduced in Section 4, obtained for various ranges of values of the scale parameter σ when using the normalized correlation as the similarity measure. The conclusion drawn from this set of tests was that increasing the σ value above 9 leads to only slight performance increases, while at the same time resulting in considerably higher computational overhead. Therefore, the range of values of the scale parameter was set to $\sigma = -9,...,9$.

σ range	$\mathbf{D_1}$ (%)	$\mathbf{D_2}$ (%)	$\mathbf{D}$ (%)
-1,...,1	12.09	12.24	12.17
-3,...,3	11.45	11.65	11.55
-5,...,5	11.13	10.76	10.95
-7,...,7	9.87	9.17	9.52
-9,...,9	**9.86**	**8.91**	**9.38**

Table 1: Comparison of tracking results for various ranges of values of the scale parameter σ with respect to error measures $D_i, i = 1,2$ (12) and D (13).

The MMTT algorithm has also been tested on a number of single-subject indoor and outdoor video sequences, as well as on studio video sequences of the reference database presented in [9]. Results of the proposed algorithm when tracking the face of the subject in one of the outdoor video sequences are illustrated in Figure 2. In this sequence, a female subject is moving, staying within the field of view throughout the sequence, with outdoors lighting conditions and no occlusion. The results clearly illustrate the successful localization of the face throughout the sequence.

Table 2 (last row) presents the results of the MMTT algorithm with respect to the error measures described earlier for all test sequences (mean value), as well as the average frame rate obtained. To enable comparison with existing template representations, the same performance measures were used to evaluate the results of the various template representations and algorithms presented in Section 3. Results are contained in rows 1-5 of Table 2. Labels "GBTT-1", "GBTT-2" and "GBTT-3" correspond to results obtained with the GBTT algorithm, when using different sets of parameter values for the Gabor wavelets. It is clear that the proposed template is superior to all others.

Algorithm	$\mathbf{D_1}$ (%)	$\mathbf{D_2}$ (%)	$\mathbf{D}$ (%)	Frame Rate
GTT-NC	13.81	13.88	13.85	5.120
GTT-SSD	15.46	15.75	15.60	6.540
GBTT-1 [7]	12.45	18.52	15.49	0.020
GBTT-2 [7]	11.23	12.64	11.94	0.080
GBTT-3 [8]	10.88	11.81	11.35	0.005
MMTT	**9.86**	**8.91**	**9.38**	2.830

Table 2: Comparison of MMTT and other template representations with respect to error measures $D_i, i = 1,2$ (12), D (13) and frame rate for all test video sequences.

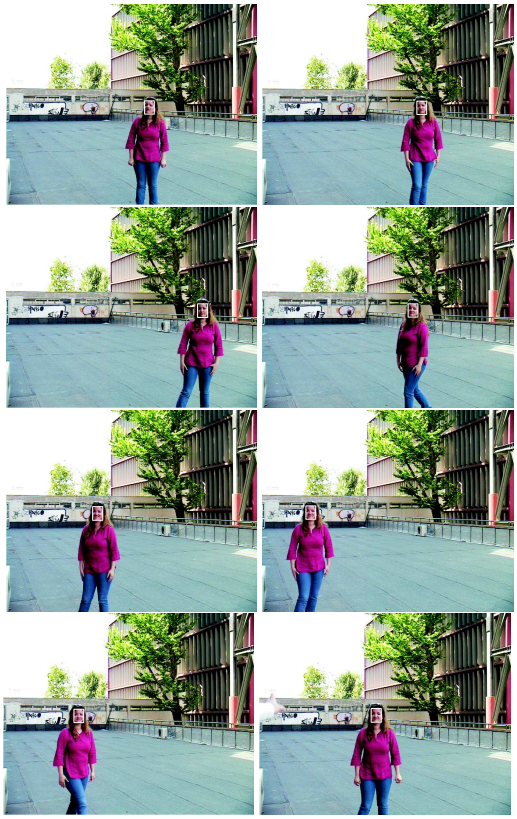

Figure 2: Tracking results of the proposed algorithm for a 350-frame segment of a video sequence. Sample frames taken at 50-frame intervals.

Figure 3 depicts a segment of a studio sequence, where the partial occlusion handling capabilities of the algorithm are illustrated. In the sequence, the face of a male subject is moving and rotating in depth, while being partially occluded by the hand of the subject.

6. CONCLUSION

In this paper, a novel template representation, the Multiscale Morphological Template, that can be used in template-based object tracking methods was introduced. The proposed template was incorporated in a template-based object tracking algorithm. Changing environment/object conditions were handled by updating the initial template over time. The algorithm was applied to face tracking with good performance in indoor and outdoor scenes, while being robust to partial occlusion. Results of the object tracking algorithm using the proposed and existing template representations were evaluated using measures based on ground truth data. The proposed template was proved to be superior to existing template representations.

7. ACKNOWLEDGEMENT

The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

REFERENCES

[1] G. Stamou, M. Krinidis, E. Loutas, N. Nikolaidis, and I. Pitas. 2D and 3D motion tracking in digital video. In Alan C. Bovik, editor, *Handbook of Image and Video Processing*. Academic Press, 2005.

[2] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

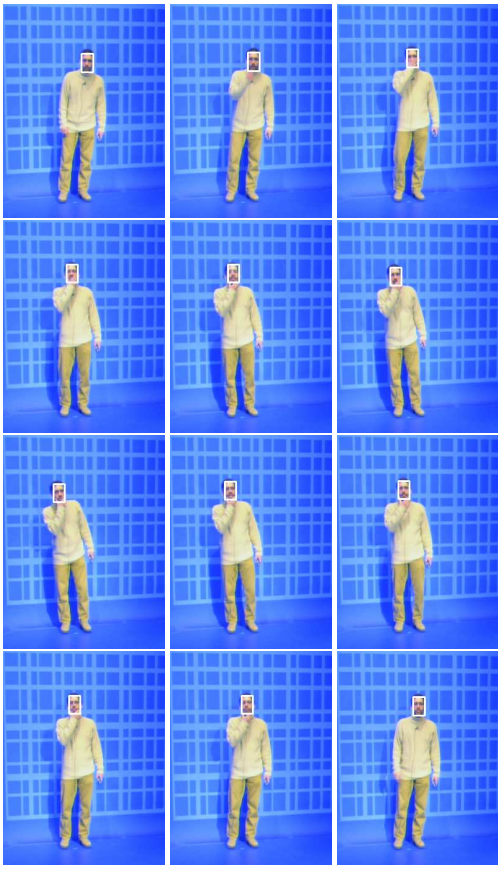

Figure 3: Tracking results of the proposed algorithm for a 120-frame segment of a studio video sequence with partial occlusion. Sample frames taken at 10-frame intervals.

[3] M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, January 1998.

[4] V. Kruger, A. Happe, and G. Sommer. Affine real-time face tracking using Gabor wavelet networks. In *International Conference on Pattern Recognition (ICPR2000)*, volume 1, pages 1127–1130, Barcelona, Spain, September 2000.

[5] P.T. Jackway and M. Deriche. Scale-space properties of the multiscale morphological dilation-erosion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):35–51, January 1996.

[6] C. Kotropoulos, A. Tefas, and I. Pitas. Frontal face authentication using morphological elastic graph matching. *IEEE Transactions on Image Processing*, 9(4):555–560, April 2000.

[7] O. Nestares, R. Navarro, J. Portilla, and A. Tabernero. Efficient spatial-domain implementation of a multiscale image representation based on gabor functions. *Journal of Electronic Imaging*, 7(1):166–173, January 1998.

[8] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.

[9] M. Krinidis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, and I. Pitas. An audio-visual database for evaluating person tracking algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005)*, volume II, pages 237–240, Philadelphia, USA, March 2005.