A COMPARISON OF FEATURES FOR REGRESSION-BASED DRIVER HEAD POSE ESTIMATION UNDER VARYING ILLUMINATION CONDITIONS

Dimitri J. Walger[1], *Toby P. Breckon*[2], *Anna Gaszczak*[3], *Thomas Popham*[3]

[1]Cranfield University, Bedfordshire, UK [2]Durham University, Durham, UK [3]Jaguar Land Rover, Warwickshire, UK

ABSTRACT

Head pose estimation provides key information about driver activity and awareness. Prior comparative studies are limited to temporally consistent illumination conditions under the assumption of brightness constancy. By contrast the illumination conditions inside a moving vehicle vary considerably with environmental conditions. In this study we present a base comparison of three features for head pose estimation, via support vector machine regression, based on Histogram of Oriented Gradient (HOG) features, Gabor filter responses and Active Shape Model (ASM) landmark features. These, reputedly illumination invariant, are presented through a common face localization framework from which we estimate driver head pose in two degrees-of-freedom and compare against a baseline approach for recovering head pose via weak perspective geometry. Evaluation is performed over a number of in-vehicle sequences, exhibiting uncontrolled illumination variation, in addition to ground truth data-sets, with controlled illumination changes, upon which we achieve a minimal ~12° and ~15° mean error in pitch and yaw respectively via ASM landmark features.

***Index Terms*—** head pose, driver head tracking, gaze tracking, pose estimation regression

1. INTRODUCTION

Knowledge of driver gaze direction provides key information about their current activity, level of alertness and general situational awareness of the road environment. As such, monitoring driver gaze through a combination of visual head pose estimation and tracking has many applications in future driver assistance [1, 2, 3, 4] and intelligent vehicle safety systems [5, 6, 7] - ranging from collision detection through to drowsiness alerting [8]. For general usage within this environment, approaches are required to be both driver invariant and robust to the highly variant illumination conditions of an in-transit vehicle interior.

From the recent survey of [9], several sensing solutions and pose recovery techniques have targeted the general problem of head pose estimation. Here we concentrate on the use of a low-cost monocular camera, offering a viable compact sensing solution for the vehicle interior, and the recovery of head orientation in terms of continuous *{pitch, yaw}* parameters. These are of primary interest for driver gaze monitoring whilst *roll* is naturally less common within this context [8].

Recent prior work in this domain has concentrated on techniques that operate directly on the image itself - commonly using edge and gradient features (e.g. [8, 10, 11, 12]). These approaches are poised against more traditional head pose estimation approaches using flexible shape modeling approaches (e.g. Active Shape Models (ASM) - Figure 1, [13]). Despite the relative age of techniques such as ASM [14] compared to contemporary techniques such as gradient distribution models [15], studies such as [13] illustrate their relevance today. To date no robust comparative study of such approaches exists grounded both over a common data-set and with relevance to the varying illumination changes of driver head pose estimation [8, 9]. An informative, over-arching survey of head pose estimation in general is provided in [9] whilst the recent notable work of [16] tackles this issue using 3D sensing via a consumer depth camera.

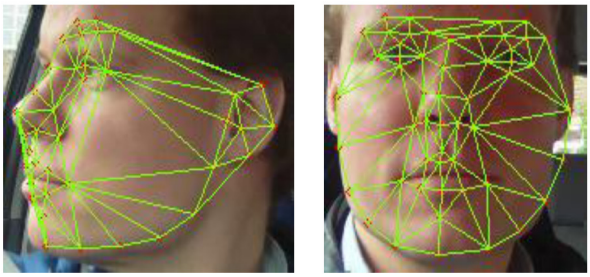

Fig. 1. Active Shape Models for frontal (left) and side profile (right) facial views.

Here we target continuous estimation of the *{pitch, yaw}* parametrization of current head pose from a monocular camera focusing on the driver head pose estimation context. Overall the most relevant prior work, within our automotive context, is that of [9] and the earlier work of [17]. [9] uses a gradient distribution approach (in essence similar to that of [18, 15]) whilst [17] compares a principle components approach to 3D motion estimation. [9] performs an extensive study of driving conditions and compares against techniques similar to those of [11] and [17]. Despite impressive results, [9] fails to address the question of evaluating varying feature inputs with a common regressive estimation framework and to address the question of relative computational performance.

In this study we concentrate on a) evaluating the performance of a range of features within a common regressive framework under varying (in-vehicle) illumination conditions, b) contrasting the relative performance of the same techniques under laboratory environment conditions using the benchmark data-sets [19, 16] and c) reporting relative run-time performance within this framework. Following an initial localization strategy akin to [9] we adopt a support

vector machine regressor approach to produce a continuous *{pitch, yaw}* output parametrization. Within this pose estimator framework we compare the use of Histogram of Oriented Gradient (HOG) features (similar to [8]), Gabor Filter responses (as per [20]) and Active Shape Model (ASM) landmark features (following the seminal work of [14]). These features are reportedly illumination invariant with respect to this task [8, 20, 14]. Our study differs significantly from prior work in the field in that it considers the in-vehicle performance of several feature types within a common regressor framework, supports this with comparative evaluation over openly available data-sets and additionally considers relative run-time performance for real-time operation.

2. POSE ESTIMATION FEATURES

We compare three head pose estimation features (Sections 2.2 - 2.4), in addition to benchmark estimation from weak perspective geometry (following [21]) as an indicator of data-set difficulty (Section 2.5), based on a common framework for initial head detection and localization within the scene (Section 2.1). Comparative results are presented in Section 3.

2.1. Detection and Localization

Face detection is initially performed using an ensemble of trained Haar-cascade classifiers [22] (a multi-orientation cascade of cascades [23, 24, 25]). Within the driver context, this ensemble of cascades is biased towards frontal profile detection (higher *a priori* probability) followed by subsequent side profile (left/right) detection (as illustrated in Fig. 3, left column). In general this offers robust initial detection within $\pm 90^o$yaw, $\pm 45^o$ pitch and $\pm 20^o$ roll offering good coverage for pose recovery (similar to [8]).

Initial detection is further integrated with a pyramidal tracking approach (Kanade-Lucas-Tomasi, [26, 27]) suitable for frame-to-frame head tracking in the fixed, although illumination varying, environment. This approach minimizes a residual pixel matching error over a large set of pixel-wise matches to recover a robust optical flow based motion estimate [26]. Here we additionally employ a forward-backward matching strategy filtering only the feature points tracked from frame t to $t+1$ that are subsequently track-able back from frame $t+1$ to t. The final motion estimation, $\hat{m}$, is computed from the optical flow of this filtered feature set via a truncated mean. Inter-feature distances, $d^t_{(i,j)}$, are then computed for each (forward-backward) matched feature pairing, (i, j), in each frame t from which a feature-wise inter-frame scaling factor, $s_{(i,j)} = \frac{d^{t+1}_{(i,j)}}{d^t_{(i,j)}}$, is computed. A mean frame to frame scaling factor, $\hat{s}$, is calculated as the truncated mean of this set. This combined spatial transformation, $(\hat{m}, \hat{s})$ facilitates onward frame-to-frame tracking of the facial region initially identified via initial detection with explicit facial re-detection presiding over a tracker prediction for future in-frame localization. Despite underlying assumptions of spatial coherence (i.e. consistent local optical flow motion) which are readily broken by head rotation or partial occlusion (e.g. hands over/on face) this provides a quantitative 91-94% correct facial region localization against publicly available data-sets [19, 16]. Overall, this extends the detection strategy of [8] within the context of driver head pose estimation.

2.2. Histogram of Oriented Gradient

From our localized face region (Section 2.1), we first use the Histogram of Oriented Gradient (HOG) of [15] features following a gradient approach akin to that of [8]. The HOG descriptor is based on histograms of oriented gradient responses in a local region around a given pixel of interest. Here a rectangular block, pixel dimension $b \times b$, is divided into $n \times n$ (sub-)cells and for each cell a histogram of gradient orientation is computed (quantised into H histogram bins for each cell, weighted by gradient magnitude). The histograms for all cells are then concatenated and normalised to represent the HOG descriptor for a given block (i.e. associated pixel location). For image gradient computation centred gradient filters $[-1, 0, 1]$ and $[-1, 0, 1]^T$ are used as per [15].

By re-sampling the localized facial region to a 64×64 pixel image, we then compute the global HOG descriptor of this localized region using a block stride, $s = 8$ ($H = 9, n = 4, b = 16$), to form the input to a Support Vector Machine (SVM) regressor. Four SVM regressors are trained, one for each of frontal and side profile facial view predicting either *{pitch, yaw}* respectively (following [8]). Based on this 1764 dimension input (i.e. $H \times n \times (s-1)^2$) we use a Radial Basis Function (RBF) kernel, with grid-based kernel parameter optimization, within a cross-validation based training regime. Training is performed over$\sim$700 example images from [16] sub-divided into frontal profile (i.e. $\{pitch, yaw\} = \{\pm 60, \pm 45\}$) and side profile (i.e. $\{pitch, yaw\} = \{\pm 60, \pm(-90 \rightarrow -45)\}$). A single side profile regressor is trained over which both left and right side profiles are evaluated via a symmetrical pre-transformation (based on profile detection, Section 2.1).

2.3. Gabor Filters

Secondly we consider the use of multiple Gabor filter response features over the same localized face region (Section 2.1). Gabor features are widely used to extract information from images [28]. In order to extract the Gabor feature information $r(x, y)$, $(x, y) \in \Omega$, we convolve the image $I(x, y)$ with the Gabor filter function $g(x, y)$ as follows:-

$$r(x, y) = \iint_{\Omega} I(\xi, \eta) \times g(x - \xi, y - \eta) \, \mathrm{d}\, x \, \mathrm{d}\, y \qquad (1)$$

The two dimensional Gabor filter is defined as the product of two functions:- the carrier, $s(x, y\phi, \theta)$, a complex sinusoid of spatial frequency ϕ, with orientation θ, and the envelope, $w_r(x, y, \sigma)$, a Gaussian kernel of width σ as follows:-

$$g(x, y) = s(x, y\phi, \theta) w_r(x, y, \sigma) \qquad (2)$$

The carrier determines the wavelength (the preferred spatial frequency of analysis). Here we use the response of the

Gabor filter to locally characterize the face region as a summary vector feature. Following the in-depth analysis of [29] we use a single spatial frequency, $\phi = \{1.82\}$, with a set of four orientations $\theta = \{0, -45, -90, -135\}^o$ resulting in four magnitude response values per input pixel over the face region (re-sized to a 20×20 pixel image). This 1600 dimensional vector (i.e. $20 \times 20 \times |\{\theta\}|$) forms the input to a multiple SVM regressor approach following the same approach outlined in Section 2.2.

2.4. Active Shape Model Landmarks

Within the localized facial region, here an Active Shape Model (ASM) [14] is used to perform sub-facial feature localization. Essentially, each face within a given training set is represented as a set of landmark points corresponding to explicit facial features (see Fig. 1) over which a Point Distribution Model (PDM) is constructed via Principle Components Analysis (PCA). In operation, ASM exploit a linear formulation of the PDM by performing an iterative search to fit the model to a new unseen (facial) image post-training. Here, by priming the ASM with the output region from the tracker we localize this search within the image. As the ASM is reliant on all of the facial landmarks being un-occluded within the image we train separate ASM models for frontal and side profile respectively [14] (Fig. 1). These ASM are constructed using 53 facial landmarks (frontal) and 39 facial landmarks (profile) over $\sim$ 700 example images from the data-set of [19] following the approach of [14].

An input to head pose estimation is formed as the fully-connected set of inter-landmark distances and landmark positions normalized by the size of the localized face region (from Section 2.1). These forms a set of $v \times (v-1)$ distances and v 2D positions for the vertices of the fully-connected ASM graph corresponding to the facial landmarks ($v \in \{v_{frontal}, v_{side}\} = \{53, 39\}$) resulting in an $(v^2 + v)$ dimensional representation (i.e. $(v \times (v-1)) + 2v$) for each ASM respectively (Fig. 1). Subsequently, this forms the input to a multiple SVM regressor approach following the same methodology outlined in Section 2.2 for the prediction of *{pitch, yaw}*.

2.5. Weak Perspective Geometry

Finally we consider the work of [21] as a baseline technique in this study to provide an accessible measure of data-set/scenario difficulty. This uses a simple geometric approach for face pose estimation based on knowledge of a semantic set of five facial features, $f_{face} = \{$*tip of nose, extremities of the eyes (x 2), extremities of mouth*$\}$ (see Fig. 2, white). Based on localization of these features a line is drawn between the midpoint of the eye and midpoint of the mouth (Fig. 2, red). A statistical ratio thus facilitates recovery of the base of the nose [21] from which the vector *(nose base* $\rightarrow$*nose tip)* can be estimated as a 2D projection of the facial normal (Fig. 2, green). Estimation of the projective transformation from the standard spatial layout of f_{face} to that of the example facilitates direct recovery of a *{pitch, yaw}* estimation [30]. Here

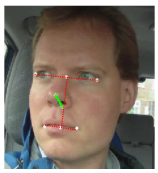

Fig. 2. Facial normal estimation (green) from facial feature geometry (white/red).

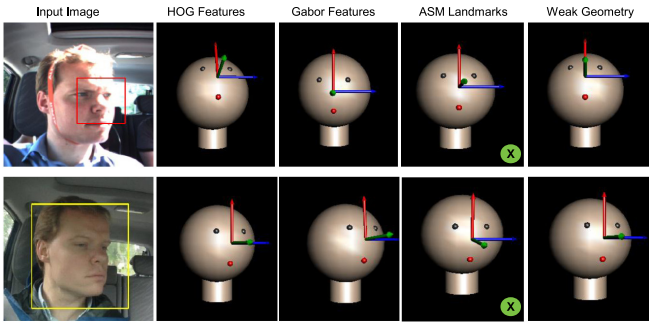

Fig. 3. Head pose estimation under varying illumination conditions (upper and lower)

localization is based on a combination of ASM, {*extremities of the eyes (x 2), extremities of mouth*}, and a specific trained Haar-cascade classifier [22], {*tip of nose*}, which empirically offered superior localization to that of the ASM for this feature.

Inclusion of this approach [21] makes the results of this study more directly comparable to those of other studies [9] and provides a baseline for comparison of the other features within the common regression framework.

3. EVALUATION

Our comparative evaluation is based on qualitative and quantitative analysis (Sections 3.1, 3.2) and analysis of relative computational performance (Section 3.3).

3.1. Qualitative Results

Our qualitative evaluation is based on an in-vehicle data-set captured from a centrally mounted camera facing the driver (e.g. Fig. 2). This data-set is captured over multiple circuits of a given route at varying times of day ensuring varying illumination conditions within any given circuit (as vehicle position changes relative to the sun) and between circuits due to changes in ambient illumination conditions. The route comprises of a complex (campus) site involving numerous illumination occluders, sources of shadow into the vehicle and periodic shadow to bright illumination changes (caused by trees, street furniture, building occlusion etc.). It is notably difficult to quantify illumination non-uniformity under such operating conditions.

In Fig. 3 we initially illustrate the comparative performance of the four different approaches (Sections 2.2 - 2.5) under varying lighting conditions denoting the most plausible pose in each case (green "x"). From this figure we can see that the use of ASM facial landmarks generally outperform the other approaches recovering both the (driver) left

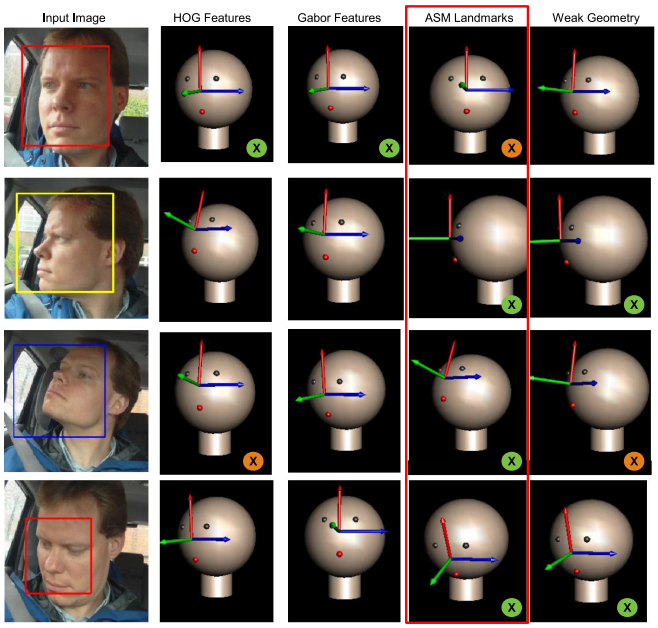

Fig. 4. Head pose estimation under varying non-uniform illumination conditions

head yaw with minimal variation in pitch from the horizontal (Fig. 3 upper) and (driver) left head yaw with mild downward pitch/glance (Fig. 3 lower). ASM landmarks appears to be the only approach capable of recovering such subtleties correctly.

These qualitative results are further illustrated in Fig. 4 where we see a range of varying illumination head pose examples from this test data-set. The comparative performance of each technique is shown with the plausible poses (green "x") and semi-plausible poses (orange "x") identified. From this illustrative example we can see that the ASM landmarks approach consistently produces a plausible pose estimation followed by the weak geometric estimation approach (Fig. 4). The HOG and Gabor approaches often produce implausible pose estimates (e.g. Fig. 4, rows 2 and 4.). Fig. 3 and 4 are representative of the results obtained over the entire test sequences. The bounding box (left) represents the facial region localized from detection/tracking forming an input to the specific feature approach (Fig. 4/3, red = frontal profile detection, yellow = side profile detection, blue = tracked position).

The Gabor and HOG approaches appear to be less robust to variation in illumination (Fig. 3 and 4) and head localization error (e.g. Fig. 3, row 1 and Fig. 3, row 3). In general, ASM landmarks were observed to be more robust both to head localization error (Section 2.1), changes in illumination and to mild partial occlusion. This is potentially attributable to the fact that the local normalization step in most gradient distribution approaches (e.g. HOG or [8]) assume locally consistent illumination, which may vary globally from instance to instance, rather than the extreme illumination gradient obtained over some of the test examples (e.g. Fig. 3, row 1 (left)). The magnitude of the Gabor filter response (as per [20]) suffer from a similar trait, as will all approaches inherently based on having a local gradient magnitude profile that is consistent over the localized facial region in any given example (e.g. [8, 10, 11, 12]).

Data-set	Localization	HOG	Gabor	**ASM**	Geometry
A [19]	91%	y: 29.2	y: 22.7	**y: 20.9**	y: 25.6
		p: 25.7	p: 32.6	**p: 19.6**	p: 24.8
B [16]	94%	y: 15.1	y: 15.8	**y: 14.6**	y: 18.6
		p: 13.2	p: 12.6	**p: 11.9**	p: 13.7

Table 1. Mean absolute error in *{pitch (p), yaw (y)}* (degrees)

	Localization	HOG	Gabor	**ASM**	Geometry
$ms.$	33	40	96	**21**	50

Table 2. Mean Execution Time per Image ($ms.$)

3.2. Quantitative Results

We further support our qualitative observational study with a quantitative comparison using the public data-sets described within [19] and [16] for which ground truth *{pitch, yaw}* is available (Table 1). From Table 1, whilst we can generally observe a lower mean absolute error over data-set B ([16]) than A ([19]). Whilst B achieves higher successful face localization and lower weak geometry based pose error (indicating relative difficultly), the error achieved by ASM is consistently lower than the other techniques against ground truth and is comparable to those achieved using gradient distribution approaches in [8] and others [10, 11, 12, 9]. Data-sets A and B use controlled illumination variations.

3.3. Computational Performance

Within the driver head pose estimation context, considering aspects of the real-time performance, we additionally present execution time per image frame (in *ms.*) in Table 2 (hardware platform: Intel core i5 CPU (3M Cache, 2.26 GHz), Windows 7 64-bit). With a standard localization overhead of 33*ms* for face localization, we can observe a significant computational advantage of the ASM landmarks approach over the other technique in terms of achievable frame-rate (approx. 18 *fps*) against the slowest (Gabor, approx. 8 *fps*).

4. CONCLUSIONS

Overall, we conclude that the use of ASM landmark features [14] outperform contemporary gradient distribution (e.g. [8]) and Gabor filter response type (e.g. [20]) features as an input to regression based estimation of head pose under varying illumination conditions. Whilst these techniques do perform moderately well, we show that ASM landmarks qualitatively perform better in the recovery of the more subtle aspects of pose under complex and varying illumination conditions - as commonly found in driver in-vehicle head pose estimation. This is quantitatively supported by lower mean absolute error in *{pitch, yaw}* estimation over established data-sets [19, 16]. Furthermore, ASM landmarks [14] offer significantly greater computational efficiency than contemporary gradient distribution approaches (e.g. [8, 18, 15]). Future work will expand this study following the evaluation methodology of [16].

5. REFERENCES

[1] M. L. Eichner and T. P. Breckon, "Real-time video analysis for vehicle lights detection using temporal information," *Proc. 4th European Conference on Visual Media Production*, p. 15, 2007.

[2] ——, "Integrated Speed Limit Detection and Recognition from Real-Time Video," in *Proc. Intelligent Vehicles Symposium*. IEEE, May 2008, pp. 626–631.

[3] I. Tang and T. P. Breckon, "Automatic Road Environment Classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 476–484, Jun. 2011.

[4] A. Kheyrollahi and T. P. Breckon, "Automatic Real-time Road Marking Recognition Using a Feature Driven Approach," *Machine Vision and Applications*, vol. 23, no. 1, pp. 123–133, 2012.

[5] F. Mroz and T. P. Breckon, "An Empirical Comparison of Real-time Dense Stereo Approaches for use in the Automotive Environment," *EURASIP Journal on Image and Video Processing*, vol. 13, 2012.

[6] A. M. Heras, T. P. Breckon, and M. Tirovic, "Video Resampling and Content Re-targeting for Realistic Driving Incident Simulation," in *Proc. 8th European Conference on Visual Media Production*, Nov. 2011, pp. sp–2.

[7] O. Hamilton, T. Breckon, X. Bai, and S. Kamata, "A Foreground Object based Quantitative Assessment of Dense Stereo Approaches for use in Automotive Environments," in *Proc. International Conference on Image Processing*. IEEE, 2013, pp. 418–422.

[8] E. Murphy-Chutorian, A. Doshi, and M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *Proc. Intelligent Transportation Systems Conference*, 2007, pp. 709–714.

[9] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[10] Y. Li, S. Wang, and X. Ding, "Person-independent head pose estimation based on random forest regression," in *Int. Conf. Image Processing*, 2010, pp. 1521–1524.

[11] M. Shafi, F. Iqbal, and I. Ali, "Face pose estimation using distance transform and normalized cross-correlation," in *Proc. Conf. Signal and Image Processing Applications*, 2011, pp. 186–191.

[12] H. Ho and R. Chellappa, "Automatic head pose estimation using randomly projected dense sift descriptor," in *Proc. Int. Conf. Image Processing*, 2012.

[13] M. Jiang, L. Deng, L. Zhang, J. Tang, and C. Fan, "Head pose estimation based on active shape model and relevant vector machine," in *Trans. Int. Conf. Systems, Man, and Cybernetics*, 2012, pp. 1035–1038.

[14] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[15] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 886–893.

[16] G. Fanelli, J. Gall, and L. van Gool, "Real time head pose estimation with random regression forests," in *Proc. 33rd Annual Symposium of the German Association for Pattern Recognition*, 2011, pp. 617–624.

[17] Y. Zhu and K. Fujimura, "Head pose estimation for driver monitoring," in *Proc. Intelligent Vehicles Symposium*, 2004, pp. 501–506.

[18] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[19] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. International Workshop on Visual Observation of Deictic Gestures (ICPR)*, 2004, pp. 1–9.

[20] V. Krüger and G. Sommer, "Gabor wavelet networks for efficient head pose estimation," *Image and Vision Computing*, vol. 20, no. 9, pp. 665–672, 2002.

[21] A. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, no. 10, pp. 639–647, 1994.

[22] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[23] T. P. Breckon, S. E. Barnes, M. L. Eichner, and K. Wahren, "Autonomous Real-time Vehicle Detection from a Medium-Level UAV," in *Proc. 24th International Unmanned Air Vehicle Systems Conference*, Mar. 2009, pp. 29.1–29.9.

[24] A. Gaszczak, T. P. Breckon, and J. W. Han, "Real-time people and vehicle detection from UAV imagery," in *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, Jan. 2011, p. Vol. 7878 Number 78780B.

[25] T. Breckon, A. Gaszczak, J. Han, M. Eichner, and S. Barnes, "Multi-Modal Target Detection for Autonomous Wide Area Search and Surveillance," in *Proc. SPIE Emerging Technologies in Security and Defence: Unmanned Sensor Systems*. SPIE, 2013, pp. 1–19.

[26] J. Shi and C. Tomasi, "Good features to track," in *Proc. Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[27] J. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, 2001.

[28] S. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on gabor filters," *IEEE Trans. Image Processing*, vol. 11, no. 10, pp. 1160–1167, 2002.

[29] B. Gokberk, L. Akarun, and E. Alpaydin, "Feature selection for pose invariant face recognition," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 306–309.

[30] C. J. Solomon and T. P. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010.