Real-time occupant detection system in an active illumination environment

J.J. Yoon and T.J. Ellis[1]
City University, London, EC1V OHB, UK
[1]Kingston University, Surrey, KT1 2EE, UK

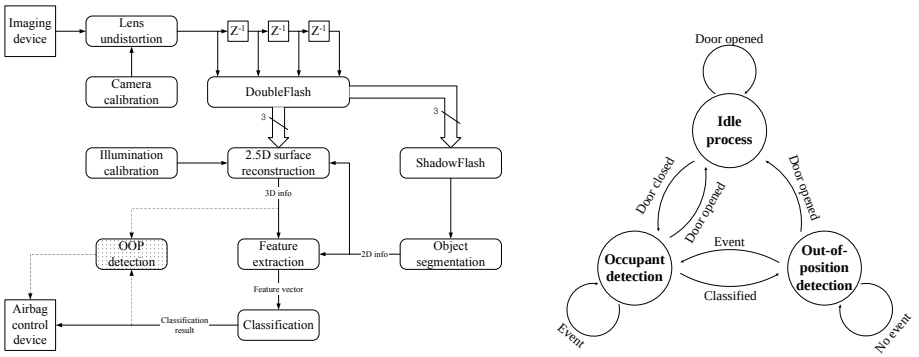
Abstract

A single grey-scale camera based object classification system for vehicle airbag deployment control in wide and frequent illumination variations is introduced. Image sequences are acquired using an active illumination systems that is used to minimise the effects of the widely varying levels of ambient illumination, combined with a means of shadow suppression. Two-dimensional information of the object is extracted by employing the active contour model, based on *a priori* knowledge of the passenger behavior. A triplet of images, of which each image is illuminated from a different direction, are sequentially used by the photometric stereo method to recover the three-dimensional shape of the object. Utilizing both the two and three-dimensional properties of the object, a 29-dimensional feature vector is defined for the training of a neural network designed to solve a three-class problem, with the classes being *forward-facing child seat*, *rear-facing child seat*, and *adult*. The system is tested on a database of over 84,000 frames collected from a wide range of objects in various illumination conditions. A classification accuracy of 98.9% was achieved within the decision-time limit of three seconds.

1 Introduction

According to the Federal Motor Vehicle Safety Standard (FMVSS) 208 set out by U.S National Highway Transportation and Safety Administration (NHSTA), nearly 100 percent of all automobiles sold in US must have the ability to automatically control the deploying power of airbags based on crash severity, occupant type and size, as well as seat belt usage, starting with the 2006 model year [8]. As manufacturers began to develop various *occupant detection systems*, the vision techniques have attracted much attention due to their superior adaptability to various vehicle cabin environments as compared to the other mechatronic methods. In recent years, a number of optical approaches have been studied to resolve the airbag suppression decision problem [5, 7]. These studies can be classified into two categories depending on the number of cameras used in the system. In the earlier versions of occupant detection systems, single camera approaches were in demand due to the high cost of imaging sensors. However, such monocular systems did not provide sufficient 3D information necessary for functions such as the *out-of-position detection*, which is a supplementary task guaranteeing low risk deployment according to the position/pose of the passenger. As a consequence, the majority of occupant detection systems employ stereo vision techniques using two cameras.

BMVC 2004 doi:10.5244/C.18.52



(a) (b)

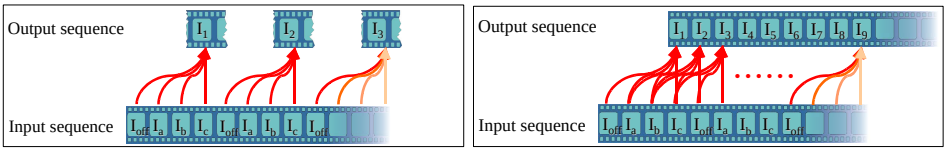
Figure 1: System overview: (a) the structure of the proposed system in conjunction with the out-of-position detection system, and (b) the state transition diagram of the overall system. The transition 'Event' occurs when any dramatic change happens in the field of view, such as any abrupt change of classes.

Faced with the increasing demand for various vision-based in-vehicle applications, the growing number of cameras employed has come under serious scrutiny. For this reason, this research focused on developing a single camera system able to generate additional 3D information by using minimal supplementary active illuminations. The aim of this paper is to propose a novel framework mainly for, though not restricted to, the occupant detection system, as well as to demonstrate the possibility of alternative systems with comparable performance to binocular based vision systems. The proposed system is designed to classify an object in a vehicle for facilitating the airbag control module. Fig.1(a) shows a basic framework of the system. It is assumed that the classification results may be shared with the out-of-position (OOP) detection system as shown in Fig.1(b) describing the state transition between two systems. The OOP detection is activated only if the object is classified as an *adult*, which is the only class continuously observed after the classification.

2 Image acquisition and pre-processing

2.1 Illumination stabilisation: DoubleFlash

Mainstream CCD based, and most of the emerging CMOS based image sensors, do not provide sufficient optical dynamic range for monitoring the interior of a vehicle which are subject to extreme variations of illumination both spatially and temporally [5]. In order to capture images without losing image details in such an environment, it is essential to employ an imager with a high dynamic range and/or a novel approach to decrease the dynamic range without varying illumination offset. The DoubleFlash technique was employed in the proposed system, which combines the advantages of offset reduction and dynamic range compression by illuminating two input images with different radiant intensities, originally introduced in [6].



(a) (b)

Figure 2: Real-time ShadowFlash: (a) without the sliding N-tuple strategy (b) with the sliding N-tuple strategy.

2.2 Shadow removal: ShadowFlash

Nearly all vehicle interior monitoring applications introduce supplementary light sources (usually in the near-infrared region) in order to attain an appropriate illumination offset. Therefore, strong cast shadows are unavoidable in the field of view. Shadows often generate erroneous segmentations causing false detection of imaginary objects, which hinders the overall performance of a system. The ShadowFlash introduced in [9] is a method to eliminate shadows by simulating a virtual light source of infinite size. The algorithm uses multiple images, where each image has been flashed from a different direction. The number of input images N_{in} necessary to create one shadow-free image is equal to the number of employed light sources n_{light} plus an additional image for calculating the ambient light suppression.

$$N_{in} = n_{light} + 1 \tag{1}$$

The experiments are performed with *three light sources* (the minimum number for a practical photometric stereo method), making the number of inputs four, including ambient lighting. If the ambient illumination image I_{offset} is negligible, the number of input images can be reduced to n_{light} by discarding the DoubleFlash. However, the robustness to deal with illumination change is lost.

2.3 Temporal domain processing: sliding n-tuple strategy

The ShadowFlash idea can be extended to the temporal domain by synchronising the illumination sources with the trigger signal of a imager so that the imager produces a video sequence of $(\cdots, I_b, I_{offset}, I_a, I_b, I_{offset}, I_a, \cdots)$ where I_x are the images illuminated by the light source x while I_{offset} represents an image having only ambient illumination. However, the direct application of the ShadowFlash method to the temporal domain raises two problems. First, the frame rate of the output sequence will be reduced to $\frac{1}{N_{in}}$ accompanied with a n_{light}-frame delay in the beginning of the acquisition, because N_{in} images are required to obtain one shadowless image as explained in Eqn.1. Secondly, if any object in the scene moves during a N_{in}-tuple, some artifacts will occur around the boundary of the object.

In order to avoid the frame rate reduction, a *sliding N-tuple strategy* is proposed. A memory window with the width of N_{in} frames is created, whereby the window is moving along the time axis. In the window, N_{in} differently illuminated successive images are constantly refreshed. These images continuously form a set of inputs to create a shadow-free output image. Fig.2(a) shows that the frame rate of the result sequence is divided by four while the output frames are consecutively calculated by employing the sliding

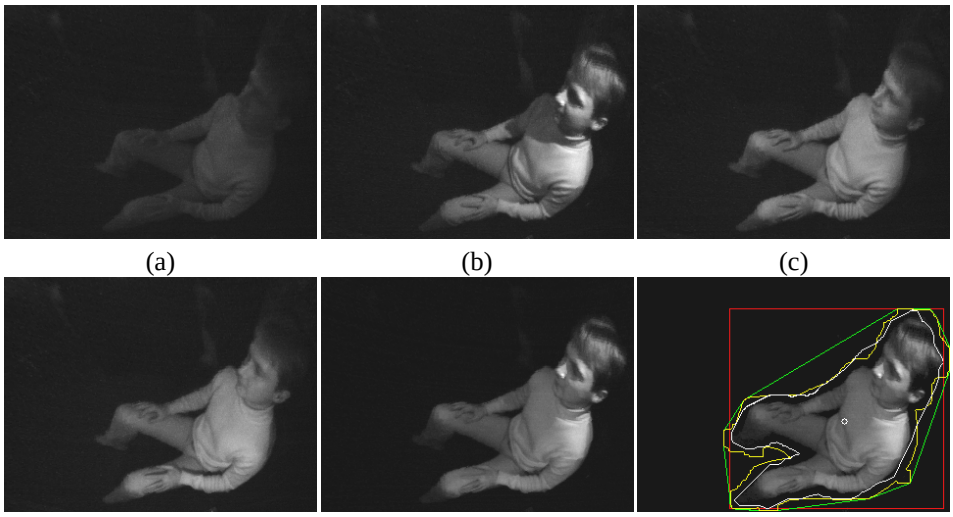


Figure 3: ShadowFlash and segmentation results: (a)-(d) a sequence of images used for this test: I_{offset}, I_a, I_b and I_c, respectively, (e) the ShadowFlash image and (f) segmentation result applied to the ShadowFlash image. (red: bounding box, green: convex hull, yellow: approximate boundary and white: snake result)

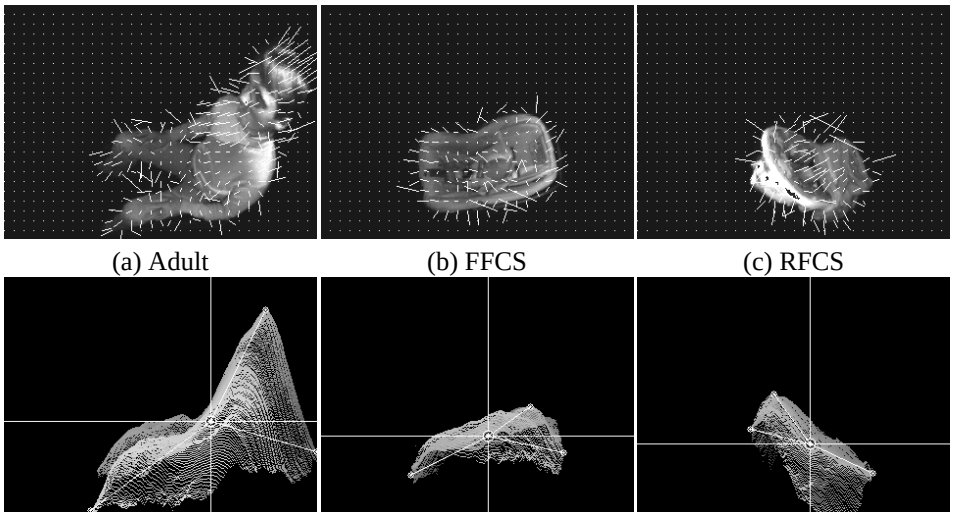
N-tuple strategy in Fig.2(b). Fast moving objects may distort the result of the sliding N-tuple strategy. The amount of distortion depends on the frame rate of the imager. When the imager produces frames with sufficient speed, the artifacts caused by moving objects should be negligible. In case of a slow frame rate compared to the velocity of moving objects within the scene, a supplementary algorithm should be implemented to detect and correct the difference between frames. However, if such a correction filter is added to the ShadowFlash approach, the speed advantage over the other algorithms will be reduced or lost. An example of the extended ShadowFlash method is shown in Fig.3(e).

3 Extracting information from 2D and 3D processing

3.1 Object boundary extraction: active contour models

Like many other machine vision applications, boundary extraction is of great importance in the proposed system to provide object outline shape information. Once the segmentation process starts, the textural similarity of each frame against to the *reference background* is analysed by comparing their local statistics by window operation (5×5). Since the local and global illumination changes are stabilised by the DoubleFlash, and all the shadows are removed by ShadowFlash, the statistics comparison followed by a simple *adaptive thresholding* is sufficient to provide an *approximate boundary* of the observed object.

An *active contour model* [3] is employed to refine this approximate boundary. In order to provide an *initial contour* for the following snake evolution, a *convex hull* is generated



(a) Adult (b) FFCS (c) RFCS

(d) Adult (e) FFCS (f) RFCS

Figure 4: The surface normals and their recovered 3D surfaces projected along the *z*-axis: (a)-(c) needle maps superimposed on their original images and (d)-(f) surface reconstruction results with the spread axes.

around the approximate boundary. In this case, a sequence of the approximate boundary points exists normally between two consecutive convex hull vertices, while each pair of vertices form a line segment. For each sequence, we can define *convexity defect* as the maximum vertical distance between the sequence and corresponding line segment. For example, the convexity defect of a sequence adjacent/overlapped to the corresponding line segment is zero. Finally the convexity defect is used for weighting the energy function of each snake cell which belongs to the line segment providing the weight, so that the cell has higher mobility when it has a greater distance to the approximate boundary than other cells. A segmentation result is shown in Fig.3(f).

3.2 Three dimensional vision: the photometric stereo method

Since the goal was to provide three dimensional information without using a binocular imager, there are three main possibilities to consider for surface recovery. *Structured lighting* is a so-called active stereo vision method which calculates the three-dimensional shape of the object based on the deformation of the light patterns projected on the target's surface. The calculations are simple and fast so that the shape of the scene could easily be extracted, provided that the feature points of the projected pattern are accurately detected. However, in reality, it is difficult to implement an accurate pattern using an infrared light source due to the constant vibration in the vehicle environment. Furthermore, such patterns may not provide enough resolution for object classification.

Recently, a *time-of-flight* (TOF) imager, which consists of an array of single point distance measurement units measuring the runtime or phases of the emitted light from a supplementary light source, is of great interest in the industry. The TOF imager has a great advantage in that it directly measures the absolute depth and determines a complete

distance map of the scene without any delay. Nevertheless, since the measurement range is limited by the maximum radiant power, the possibility of violating the eye safety limits still remains to be solved.

The *photometric stereo method* (PSM) is an extended version of the *shape from shading* (SFS) using multiple light sources, which constructs the relative depth of the object with *full resolution* by using its reflection properties. Unlike the SFS, which suffers from the lack of sufficient information in an arbitrary irradiance image to reconstruct the object surface unambiguously, it was successfully proven that the PSM performs the surface recovery with greater ease, especially when there are more than three light sources.

Since the multiple illuminations are already employed for the ShadowFlash method, it is possible to apply the PSM, for there was no need to provide supplementary hardware for such an implementation. The problem of using the PSM method for our application was any abrupt movements of objects in-between two successive frames which may cause significant distortion of the recovered surface. However, after extensive testing, it was concluded that the level of distortion caused by motion is acceptable for the application considered here, which do not need to make a decision frame-wise, and especially for systems which do not require high spatial resolution of the scene. The frame rate of the imager is also a primary factor which influences the reconstruction performance.

The overall task of the PSM involves two major procedures: estimation of surface normals, and integration of the object surface from the normal vectors. The estimation of the surface normal vector could be performed *albedo-independently* by solving irradiance equations supported by *a priori* information about the direction and power of the illumination sources [4]. The Frankot-Chellappa algorithm [1], based on minimising integrability conditions in the frequency domain, is employed with a minor modification to improve its robustness for small artifacts caused by motion regardless of its disadvantage in computation time. Some typical surface recovery examples and needle maps of their surface normals are shown in Fig.4.

4 Classification

4.1 Feature selection

Because there is no need for detecting an empty seat for a safety reason, the number of occupant types to be classified is limited to three: *adult*, *forward-facing child seat*(FFCS) and *rear-facing child seat*(RFCS). Although seeking *distinguishing features* invariant to any irrelevant transformations of input is an essential task to make the job of the classifier trivial, it was still difficult to find apparent features which clearly discriminate all three classes. Therefore, each feature is designed to specify at least one class from the other two (e.g. use of the occupant size in order to distinguish an adult from the child seat classes). The proposed 29 features are defined as follows:

Extended Gaussian image: 4 dimensions The EGI is a histogram of the surface normals computed over a discretised Gaussian sphere. While the surface normals are easily derived during the calculation of the PSM (see Fig.4), it is expected that the rear-facing child seat should have a different aspect of its surface direction from the ones from the other two classes. The histogram is divided into bins of 90 degree each, and the number of the normal vectors belongs to each bin are calculated.

Surface depth: 4 dimensions The profile of the relative depth projected from the top of an object is also used as a feature. Since the camera coordinate system differs from the world

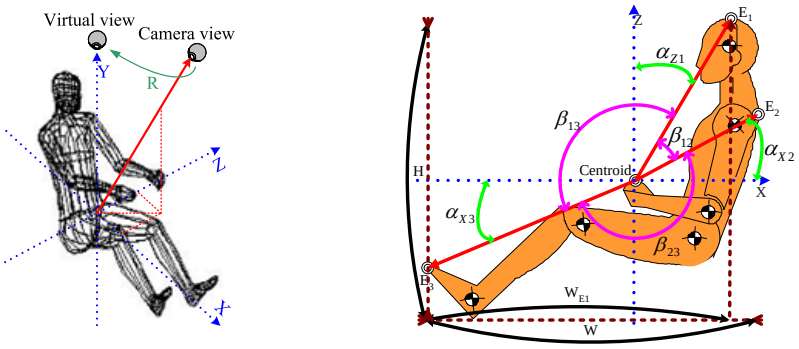


Figure 5: Feature extraction:(a) The camera calibration provides the rotational matrix **R** with respect to the world coordinate origin. In principle, all the three-dimensional features are rotationally transformed in order to make them correctly viewed from the top. (b) The extrema E_1, E_2 and E_3 are defined as a most upper, most front (left) and most rear (right) point on the recovered surface, respectively.

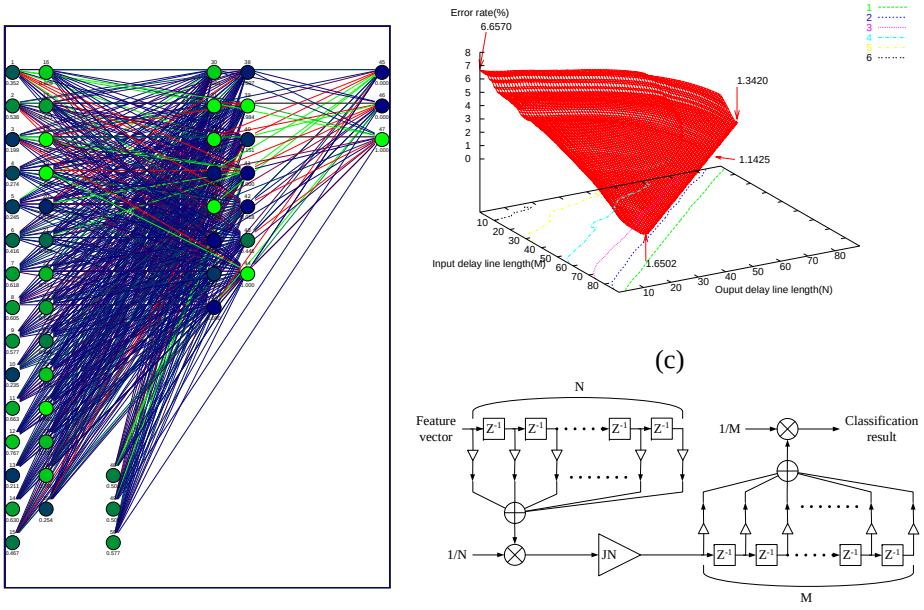
coordinate system, a rotational transformation is performed with a given rotation matrix **R** representing three partial transformation (*pan*, *tilt* and *roll angles*) in order to provide a depth profile projected along the *z*-axis of the world coordinate system. A brief illustration of changing the view point is shown in Fig.5(a).

Spread axes information: 9 dimensions With the successful recovery of the object surface, three extrema (E_1, E_2 and E_3 as defined in Fig.5(b)) on the surface are used for the definitions of a few useful features. The *spread axes* [5] are the lines between the center of gravity and the extrema. Accordingly, the *spread angles* (α_{Z1}, α_{X2} and α_{X3}) are defined as the angles between the spread axes and the coordinate system; while the *relative spread angles* (β_{12}, β_{13} and β_{23}) are the angles between the spread axes themselves. These two angle characteristics as well as the lengths of the spread axes are used as key features for the classifier. A few examples are shown in Fig.4(d)-(f).

Relative position of the upper extremum: 1 dimension The relative position of the upper extremum E_1 along the *x*-axis could be a good clue to specify the rear-facing child seat class against the other two classes. As shown in Fig.5(b), the relative position P_{E_1} is simply defined as $P_{E_1} = \frac{W_{E_1}}{W}$ where W and W_{E_1} are the width of the object and the distance along the x-axis between the $E1$ and $E3$, respectively.

Volumetric ratio and compactness: 2 dimensions Since it is not possible to recognise what happens behind the object, it is difficult to define the *volume* of the object. Even if the assumption is made that the object has a flat back side, the volume of the target may still be extremely sensitive to the segmentation result. Consequently, the ratio of the three-dimensional surface area to the two-dimensional boundary area is defined as a *volumetric ratio*, which should increase as the volume of the object expands. Assuming a flat back side, the proportion, or *compactness*, of the object volume to a hexahedron enclosing the object could also provide robust estimation of its volume.

Other 2D geometric information: 9 dimensions Three low-order components of both *normalised central moments* and *Hu moments* are selected as features, along with the *width*, *height* and *area* of the object boundary.



(a) (b) (c)

Figure 6: Classifier design: (a) the Jordan network after the learning process, (b) the proposed classifier framework with two tapped delay lines and (c) the classification error space with respect to the length of delay lines.

4.2 Classifier design

Since a change of the occupant type is not likely during driving, the classification is, in most cases, enough to be performed once in the beginning of operation unless any dramatic change in the field of view occurs (Fig.1(b)). Hence, it is assumed that the proposed system must reach a decision within *three seconds*, which implies processing of 90 frames at the 30 Hz frame rate before making a final decision.

Considering that the proposed features did not reflect any dynamic properties of the passenger, it was necessary to construct a classifier model which is able to handle and classify temporal series. Therefore, trained in a supervised way, a *partially recurrent network* proposed by Jordan [2] is employed with the support of two *tapped delay lines*, which are delay memories providing access to its contents at arbitrary intermediate delay length values. Each tapped delay line improves the accuracy of overall classification performance by filtering the noisy components in the stream of either feature vector (input) or classification result (output). The maximum delay length of the proposed system is limited to the 90 frames, allowing the system to monitor three seconds of the passenger history. The proposed Jordan network is shown in Fig.6(a) while Fig.6(b) presents the overall structure of the classifier module.

Class type	Forward facing child seat	Rear facing child seat	Adult
Error rate(%)	14.2	15.4	0.725
Favorite error	RFCS(99.5%)	FFCS(90.0%)	RFCS(68.3%)

Table 1: Error statistics *without* the tapped delay lines. Overall error rate: 6.66%

Class type	Forward facing child seat	Rear facing child seat	Adult
Error rate(%)	10.1	13.7	0
Favorite error	RFCS(100.0%)	FFCS(91.7%)	N/A

Table 2: Error statistics *with* the tapped delay lines. Overall error rate: 1.14%

5 Experimental results

The experiment was conducted with 578 image sequences collected from 29 different child seats and 25 persons with a resolution of 320×240 in 12-bit gray scale at 30 Hz in a laboratory environment which simulated actual vehicle interiors. The sequences were illuminated by three *near-infrared* LEDs which satisfied the eye-safety. The lens distortions were eliminated using the pre-calibrated camera parameters. Additional objects such as blankets and different ambient illumination conditions were used to provide diversity. By implementing the DoubleFlash and ShadowFlash techniques, the system is independent from any ambient illumination conditions, provided that the conditions satisfies the minimum requirements of the techniques discussed in [6, 9, 5]. (e.g. the irradiance of the scene must not exceed the dynamic range of the imager.)

Finally, the sequences were evenly split into two groups creating a *training* and *testing set*, while the length of the sequences varied from 100 to 500 frames depending on the occupant's behavior, and the target output values were manually surveyed. The proposed network was trained by the resilient backpropagation (Rprop) algorithm with the training set, while the regular *logistic activation* function was set to all the neurons and initial values at its synapses were randomly chosen. The learning was halted when the network reached the error minima (the mean squared output error of 0.0793 after 120 iterations).

Since the neural network only makes a single frame decision, the classification performance was evaluated with a test set according to the lengths of two tapped delay lines. Fig.6(c) shows that the system is apparently more sensitive to the length of the output delay buffer due to the recurrent network's adaptability to sequential behavior. However, as the sizes of both delay lines increased, the difference of the sensitivity became negligible. Tbl.1 shows the error analysis according to the class types without the support of the tapped delay lines. Most errors occur between the FFCS and RFCS classes due to their similar characteristics of the two-dimensional geometry, especially when the scene is altered by additional objects (e.g. a baby holding a teddy bear in the RFCS covered by a blanket). Low error rate in the adult class was achieved even with test sequences involving large amounts of motion. These are encouraging results, as the misclassification between an adult and child seat generally poses greater danger than that of the misclassification between two child seats. After applying the tapped delay lines, the error rates of all classes were dramatically decreased as shown in Tbl.2. Although the original error rate of the ordinary Jordan network reaches 6.66%, a classification rate of 98.9% was achieved after setting the lengths of the input and output delay lines to 31 and 59 respectively.

6 Conclusions

A novel frame work of classification system was introduced based on a 3D surface recovery technique using a single camera with multiple illumination sources. The necessity of the wide dynamic range under varying illumination circumstances was successfully overcome by the implementation of the DoubleFlash method. Furthermore, an extended ShadowFlash technique supported by a delay buffer was proposed to provide shadowless image sequences to a real-time vision system. The active contour model, whose energy functions were weighted based on its convexity defect property, provided boundary information of the object of interest, whereas the 3D surface was recovered by the photometric stereo method with three differently flashed images forwarded from the ShadowFlash module. Using the 29-dimensional feature vector reflecting both 2D and 3D properties of the observed object, a Jordan network with the support of two tapped delay lines was trained with supervision. Finally the system resulted in the classification rate of 98.9% within the assumed classification time-limit.

References

[1] R.T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading problem. *IEEE Trans. on PAMI*, 10(4):439–451, July 1988.

[2] M.I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proc. of the 8th annual conference of cognitive science society*, pages 531–546, Amherst, MA, 1986.

[3] M. Kass, A. Witkin, and D. Terzopolous. Snakes:Active Contour Models. In *International Conference on Computer Vision*, pages 259–268, 1987.

[4] R. Klette, K. Schluens, and A. Koschan. *Computer Vision, Three-Dimensional Data from Images*. Springer, 1998.

[5] C. Koch. *Real-time occupant detection in hich dynamic range environments*. PhD thesis, City university, October 2003.

[6] C. Koch, S. Park, T. J. Ellis, and A. Georgiadis. Illumination technique for optical dynamic range compression and offset reduction. In *British Machine Vision Conference (BMVC01)*, pages 293–302, Manchester, England, September 2001. BMVA Press.

[7] J. Krumm and G. Kirk. Video occupant detection for airbag deployment. In *IEEE Workshop on Applications of Computer Vision (WACV98)*, pages 30–35, Princeton, USA, October 1998.

[8] National Highway Transportation and Safety Administration. Federal Motor Vehicle Safety Standard # 208. 2001.

[9] J. J. Yoon, C. Koch, and T. J. Ellis. Shadowflash: an approach for shadow removal in an active illumination environment. In *British Machine Vision Conference (BMVC02)*, pages 636–645, Cardiff, Wales, August 2002. BMVA Press.