TWO METRICS OF NIGHT VISION SYSTEM PERFORMANCE

Kip Smith
Cognitive Engineering and Decision Making
USA
Jan-Erik Källhammer
Autoliv Development
Sweden
Matthias Oberländer, Werner Ritter, Roland Schweiger
Daimler AG
Germany

Paper Number 11-0360

ABSTRACT

We have developed a pair of metrics for the quantitative evaluation of the performance of pedestrian detection systems. The Metric of Similarity was designed to be used to assess how well the pedestrian-detection output of an infra-red Night Vision system matches its ground truth, that is, the relative level of fit or agreement between the locations in an image frame (measured in pixels) where the system indicates it has detected pedestrians and the locations in the frame where there actually are pedestrians. In contrast, the Metric of Salience was designed to be used to infer the level of acceptance of the system by a typical driver. These are complementary dimensions of system performance.

INTRODUCTION

The design of active safety systems is an iterative, evolutionary process. Designers continually strive to improve sensor technology, alerting software, and display design, leading to the production of new generations of commercially available systems. In response, system users (drivers, customers) become more sophisticated and demanding, providing feedback to designers and establishing a self-reinforcing cycle of system improvement.

Successive generations of systems need to be compared to ascertain not only their strengths and weaknesses but also their relative levels of driver acceptance (Källhammer, Smith, Karlsson, & Hollnagel, 2007). Designers seek to compare systems developed by different providers. The process of comparing the strengths, weaknesses, and relative levels of driver acceptance of active safety systems requires objective, replicable, and readily comprehensible metrics. This paper discusses the development of two complementary metrics designed to enable both designers and safety raters assess successive generations or alternative active safety systems.

The occasion that prompted the development of the metrics was an EU-sponsored project aimed at demonstrating the feasibility of fusing two infra-red 'Night Vision' pedestrian detections systems that use different sensor systems (European Union 7th Framework Programme, 2011). In the discussion that follows, we focus on pedestrian detection systems but mean to imply that our discussion generalizes to a wide range of active safety systems. Further, we use the verb 'detect' to mean not only that the sensor has picked up a pedestrian but also that the software and in-vehicle display have highlighted the detected pedestrian to the driver.

The role of metrics in system comparison

Figure 1 is a Venn diagram of a situation frequently faced by designers seeking to assess the relative merits of two pedestrian detection systems. System X and system Y are represented by the two large overlapping squares. The letters and symbols represent 10 pedestrian encounters. There are nine instances of pedestrian detection, seven by each system. Five pedestrians are detected by both systems but one is detected by neither. Both systems appear serviceable but in need of improvement. If designers were presented with systems X and Y, they would face the quandary of weighing the relative merits of two imperfect systems. Given the non-hypothetical nature of this quandary, designers need metrics that enable them to identify classes of events (pedestrian encounters) or incidents for which one system or the other excels. The system that performs better in more situations is likely to be preferred.

If systems X and Y were to represent successive generations of a commercial product, its designers would likely need metrics that enable them to scan large volumes of field data to indentify when, where their system failed to detect a pedestrian who should have been detected, and the relative severity of that

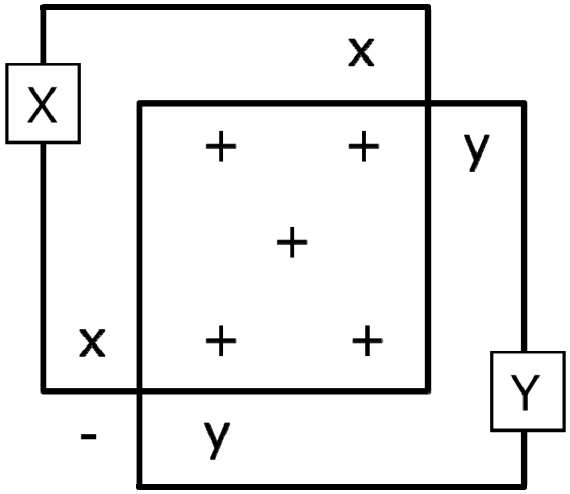

Figure 1. Venn diagram of two capable systems, system X and system Y. While both systems correctly detect most pedestrians, each misses some that the other detects. x: correct detections by system X only. y: correct detections by system Y only. +: correct detections by both systems. - : a pedestrian missed by both systems.

failure. A simple error count is not sufficient, as the severity of each error is not uniform; a system with fewer errors may have more severe failures. Further, system designers need to know whether or not drivers consider a detected pedestrian to be worthy of an alert. It does no one any good to market a system that issue alerts that drivers deem to be nuisances (Källhammer, in press).

The only time when metrics are not needed is the rare case sketched in Figure 2 in which the performance of one system dominates the other.

Data

The metrics were developed given firm constraints imposed by the nature of the data. For system X, a Far Infra-Red (FIR) pedestrian detection system, we were provided three sets of data, sequences of FIR images containing pedestrians and two sets of numerical data. The first set of numerical data was a list of the frame-by-frame coordinates of rectangles surrounding the actual locations of pedestrians in the images measured in pixels with respect to the upper left corner of the image. This data set we call the 'Ground truth', set G. The second set of numerical data was a list of the 'System output', set S, the coordinates of rectangles used by the system to highlight detected pedestrians to the driver. All entries to both numerical data sets consisted of (x,y) pairs of coordinates that contained no direct information about the distance to a pedestrian.

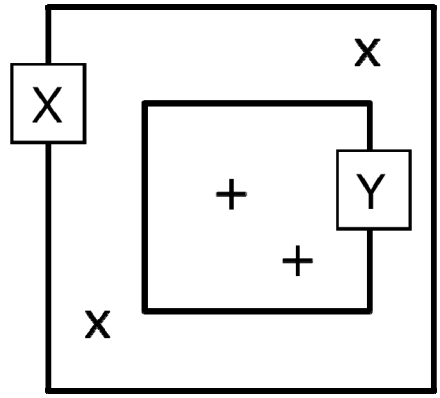

Figure 2. Venn diagram of two systems, X and Y, in which system X dominates system Y. x: corrected detections by system X only. +: correct detections by both systems.

The data constrained our task to devising quantitative metrics that define the fit of set S to set G. The degree of fit between sets affords identification of pedestrians that the system detected and those that it missed. It also affords discrimination of the similarity of the pedestrians' actual locations in the images and the locations highlighted by the system.

We were also provided a second set of system output data from a prototype system Y. These data were acquired at the same time as set G. This afforded comparison of the performance of systems X and Y.

METHOD

In this section we discuss our approach to developing the Metrics of Similarity and Salience. We begin by discussing a series of thought experiments, and a lab experiment, and their implications for the formulation of the metrics. We introduce the mathematical foundations of the metrics before turning to their formulations.

Thought experiments

The first step was to conduct thought experiments about the constraints on system performance imposed by drivers and system designers. We considered one constraint imposed by engineering concerns - the differential impact of misses and false alarms - and two constraints imposed by driver concerns - pedestrian location and proximity.

<u>Miss detections and false alarms</u> The first thought experiment addressed whether the two types of error that might be observed in the data - missed detections and false alarms - are equally important to system designers (and drivers). Figure 3 sketches our

thinking. In the upper panel, Figure 3a, a pedestrian is visible (set G) but is not highlighted by the system - there is no detection box from set S. This is a missed detection and is an error that, in certain circumstance, drivers and system designers would surely want to avoid.

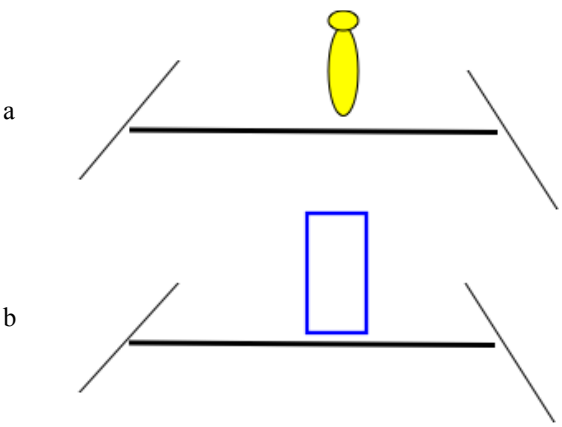

Figure 3. (A) An unhighlighted pedestrian (a miss) is worse than (B) a false alarm.

In contrast, Figure 3b shows a scene where there is no pedestrian but there is a detection box. This situation is a false alarm; the system issued an indefensible alert. Our analysis suggested that engineers will continue to refine their algorithms to suppress it (Smith, 2010).

Accordingly, this thought experiment led us to conclude that missed detections matter more than false alarms when it comes to pedestrian detection and to develop metrics that reflect this asymmetry.

Directly ahead is highly salient The second question we addressed was whether the location of the pedestrian matters to drivers (and system designers). This question has two parts. Does translation in the vertical dimension matter? Does lateral position matter? Our answers were No to the vertical dimension and Yes to lateral position.

We answered the first by finding descriptive statistics for the vertical locations of pedestrians in data set G. We found that the variance of the location of pedestrians' feet in the vertical direction was small. This means that pedestrians in our data set do not translate vertically in the images. Generally, they do not start at the top of the frame and migrate to the bottom. They usually stand or walk somewhere below the middle of the frame. We concluded that our metrics did not have to consider the vertical component of pedestrian location.

Figure 4 sketches our thinking about the lateral component of pedestrian location. In Figure 4a, a pedestrian is detected near the center of the image. In practice this means the pedestrian is more or less directly in front of the car. If the pedestrian stood still and the car continued straight, there would be a collision. This is a situation for which an alert would certainly be welcomed by drivers, system designers, and safety raters. In contrast, Figure 4b shows a pedestrian near the edge of the image. In an urban environment such a pedestrian might be walking on the sidewalk. Drivers seldom want to be alerted to pedestrians on the sidewalk. In contrast, in a rural environment, the pedestrian would likely be walking on the edge of the road, facing traffic. Drivers would likely welcome an alert to this pedestrian. This thought experiment led us to conclude that the salience of lateral location is contextually sensitive. Accordingly, we assign a greater weight to pedestrians in the center of the image than to those near the edges and retain the ability to adjust the weighting formula as a function of traffic context.

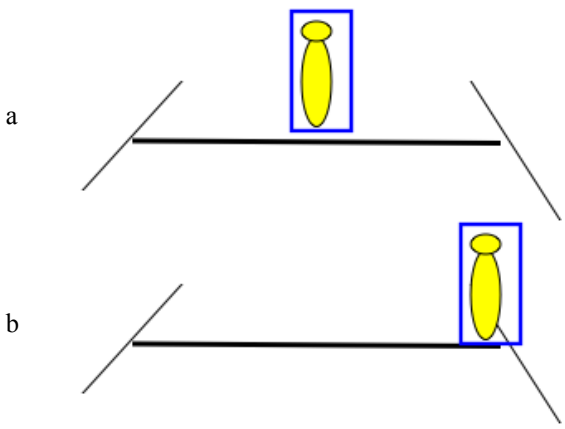

Figure 4. (A) A pedestrian in the center of the image is more salient than (B) a pedestrian on near the edge of the image.

Near is more salient than far The final thought experiment that shaped the development of the metrics concerned the proximity of pedestrians. A pedestrian who is relatively close to the car is at a greater risk of being hit by the car than a pedestrian at a greater distance. This situation is illustrated in Figure 5.

As the raw data are two dimensional projections of three dimensional space and the objects within it (e.g., pedestrians), there is no direct information about distance to pedestrians in the images. There are however two alternative approaches to inferring distance. The better method is to define the horizon and to find how far below the horizon the pedestrian is standing. This method was unavailable to us as the data sets do not contain information about the location of the horizon. The fall-back method is to

use pedestrian height as a proxy for proximity. As the car approaches, a pedestrian's apparent height increases. Both sets G and S contain information about pedestrian height.

Figure 5. Closer people pose a greater risk of collision. Closer people appear taller.

A pedestrian in the far distance is only a few pixels high. Our analysis suggested that the salience of a distant pedestrian to the driver is minimal. In contrast, there comes a time (distance) when the pedestrian becomes salient to the driver. At this ill-defined threshold, represented by the blue line in Figure 6, the pedestrian becomes a meaningful object that may influence driving behavior. Pedestrians closer than this threshold are only marginally more meaningful than they were at the threshold. These considerations suggest that the subjective mapping from height to the relative level of perceived risk is not linear. Rather, it is more likely to have a sigmoid form where the steep ramp occurs in the vicinity of the threshold distance, as sketched in Figure 6. This thought experiment led us to develop a sigmoid weighting function of pedestrian height to capture the influence of pedestrian proximity on driving behavior.

Laboratory experiment

The second step in the development of the metrics was to conduct a laboratory experiment that asked a representative sample of adult drivers to view a selected set of 15 second-long videos of pedestrian encounters recorded by the pedestrian detection system. Output (colored rectangles) from the pedestrian detection system, set S, was superposed on the videos. The participants viewed a sequence and then, individually, immediately rated the performance of the system. The procedure is discussed in detail by Källhammer & Smith (in press) and Smith and Källhammer (2010).

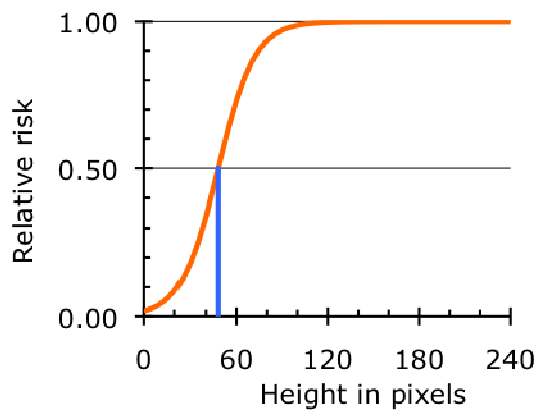

Figure 6. A sigmoid function relating height - our proxy for proximity - to relative risk.

Two findings emerged from this study. First, the participants reinforced our conclusions from the thought experiments. As expected, they were relatively unconcerned about false alarms but rated the system poorly whenever pedestrians went undetected. It appears that drivers do find missed detections more salient than false alarms. Further, the participants were less tolerant of missed detections when pedestrians stood in or crossed the road than when they stood or walked on the side of the road. We were unable to test for differential responses to proximity and distance because every pedestrian in the vido clips initially appeared in the far distance and loomed large as the vehicle drove past.

Second, participants were sensitive to both the recency and duration of the missed detection. Recency and duration are two factors long known to influence the memorability of stimuli (e.g., Baddeley & Hitch, 1993; Greene, 1986; Pavlov, 1927; Pieters & Bijmolt, 1997; Seamon, March & Brody, 1984). Recency refers to the time gap between the experience and its recall. Duration refers to the amount of time consumed by an event. For our application, recency reflects the time between (a) the last frame in the video clip in which a pedestrian was not detected and (b) the act of rating system performance for that clip. Similarly, duration is the composite time that a pedestrian went undetected in the video clip. This finding led us to conclude that recency and duration influence drivers' perception of the salience of missed detections and, hence, the relative levels at which they rate system performance.

Asymmetric distance between sets

When the system fails to detect a pedestrian, set G contains more elements than set S. Set S contains

more elements when the system posts a false alarm. The expectation of inequality in set size led us to use a MaxiMin formula to compare sets.

We calculate the distance D from one set to the other using the MaxiMin expression of Equation 1:

$$D(A,B) = \max_{a \in A}\left\{\min_{b \in B}\left[k \times d(a,b)\right]\right\} \quad (1).$$

where a and b are points in the sets A and B, respectively, and d(a, b) is the Euclidian distance between them. The free parameter k is a sigmoid weighting function that ranges from 0.0 to 1.0, like that shown in Figure 6, to map pedestrian height to the relative level of perceived risk.

When there are a different number of elements in sets A and B, D(A,B) is generally not equal to D(B,A). To appreciate this fundamental asymmetry, consider the situation sketched in Figure 7. Here there is one member of S at 10, and two of G at 12 and 17: the system finds one pedestrian but there are actually two in the image. Assuming for simplicity that k = 1, the distance D(G, S) is 7, the maximum of two values (12-10) and (17-10). In contrast, the distance D(S, G) is the maximum of the minimum of the couplet (10-12, 10-17), that is, the minimum of 2 and 7. [Euclidean distance is always positive as it is in the world.] The minimum of the couplet is 2 and the maximum of this minimum is also 2. Hence in this example the distance D(G, S) is 7 and the distance D(S, G) is 2.

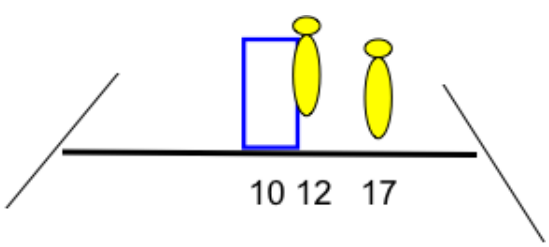

Figure 7. A hypothetical case with one system detection (set S) and two pedestrians (set G).

The important point here is that the situation shown in Figure 7 represents a miss - there are fewer elements in the system output than in the ground truth. The asymmetry of the distance calculation allows us to differentiate the effects of misses and false alarms. The calculation D(G, S) is a measure of the effect of a miss. The calculation D(S, G) is the measure of the effect of a false alarm. Here there is a miss and, accordingly, D(G, S) > D(S, G). This is a useful characteristic given that drivers and safety raters can be expected to show greater concern for misses than for false alarms (Smith, Schweiger, Ritter & Källhammer, 2011).

The Metric of Similarity

The Metric of Similarity is the normalized sum of two weighted MaxiMin distances, Equation 2. We apply two sets of weights. The free parameter $\alpha \in [0, 1]$ differentially weights misses and false alarms. For the pedestrian detection task, a miss receives the greater weight (e.g., $\alpha = 0.9$). The differential weighting emphasizes the asymmetry of the two components of the sum. The second weight k (shown in Equation 1) scales pedestrians by their height in the ground-truth image using a sigmoid function. Normalizing by the half-width of the image W/2 constrains the metric to values between 0.0 and 1.0. Because distance is a measure of difference and our goal is a metric of similarity, the normalized sum is subtracted from 1 to produce a Metric of Similarity, M.

$$M = 1 - \frac{\left[\alpha \times D(G,S) + (1-\alpha) \times D(S,G)\right]}{W/2} \quad (2).$$

The metric equals 1.0 when the system highlights every pedestrian at the same position as the ground truth. It equals 1-α in the worst case – the situation shown in Figure 3a in which an undetected pedestrian is standing directly in front of the vehicle at a distance where collision is immanent. To understand why the minimum value of the metric is 1-α, assume that the image frame shown in Figure 3a is 20 pixels wide and that the undetected pedestrian is standing directly in front of the vehicle at pixel 10. Further, in this worst case, the value of k is 1.0 because the pedestrian is near the vehicle. The value of D(G,S) is max{min[10]} and the value of D(S,G) is zero. Substituting into Equation 2 yields 1-[10 α - 0]/(20/2) which equals 1-α.

The Metric of Similarity is calculated for each frame in a sequence and plotted as function of time. An example is shown in Figure 8. If desired, the values can be summed using moving window to provide an aggregate measure of system performance per unit time.

The Metric of Salience

The Metric of Salience aims to predict the relative level of post-hoc salience of a pedestrian event to the average driver. Salience is expected to increase as the subjective experience of risk increases.

The formulation of the Metric of Salience reflects the importance of recency and duration on the memorability of failures to detect pedestrians. Equation 1 is used frame-by-frame to find the pedestrian in each frame who is associated with the

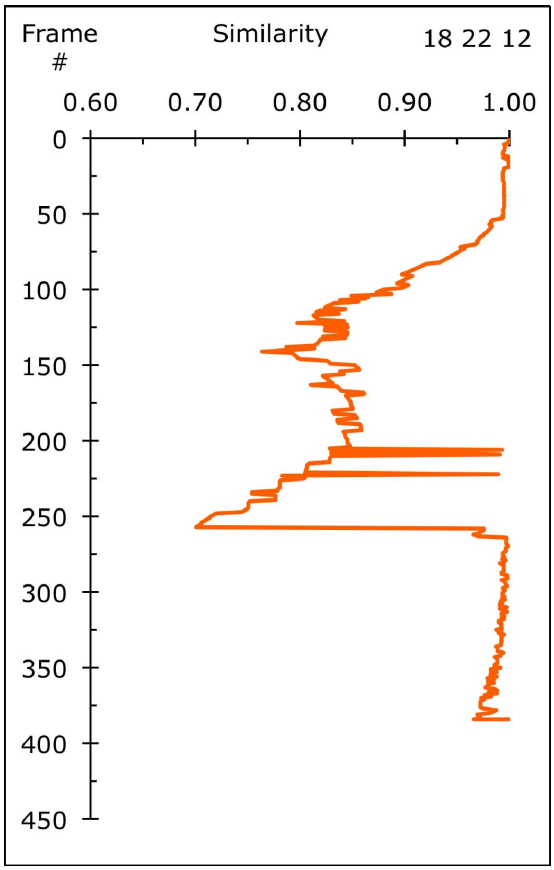

Figure 8. A time trace of the metric of similarity. Similarity, the goodness of fit of system output to the ground truth, increases to the right.

greatest distance from a system detection rectangle. We identify that pedestrian as $Max(D(G,S)_t)$ – the most salient pedestrian in the image at time t. We then find the duration of sequential frames in which a pedestrian qualifies as $Max(D(G,S)_t)$ and multiple the duration by a sigmoid function of recency that preferentially emphasizes missed detections late in the sequence of frames. The product is a single number that predicts the relative level of salience of missed detections by the system during sequence of frames.

RESULTS

We have applied the Metric of Similarity to 57 digital recordings of the output of an FIR pedestrian detection system and the corresponding ground truth data set. The sequences contain both urban and rural driving.

Low values of the metrics pointed to two opportunities for improving system performance: reducing the lag in system response and training the system to highlight pedestrians who assume odd poses. The metrics have led designers to focus on these issues as they develop the next generation of Night Vision systems with pedestrian detection.

We have also used the Metric of Similarity to scan a large data set that made it possible to compare the output from two Night Vision systems, an FIR system and a prototype system. Both systems performed well but, on occasion, failed to detect pedestrians. The metric simplified the task of identifying classes of encounters associated with missed detections. These classes were found to be essentially mutually exclusive. This result is ably represented in schematic form by Figure 1.

Figure 9 shows the match between the Metric of Salience and the average ranks of the ratings provided by participants in the laboratory study. We converted raw ratings data to ranks to correct for individual differences in scale use across participants. The lower the rank, the greater the satisfaction with the performance of the Night Vision system. Video clips that received low ranks contained undetected pedestrians that our raters expected the system to highlight. The high level of concordance among raters justifies aggregation of the ranks to calculate the average rank. The correlation between the metric salience and the average ranks of the reviewers' rating is high, r = .81. It appears that the metric predicts the relative level with which drivers are likely to be displeased when a system fails to issue an alert to an at-risk pedestrian.

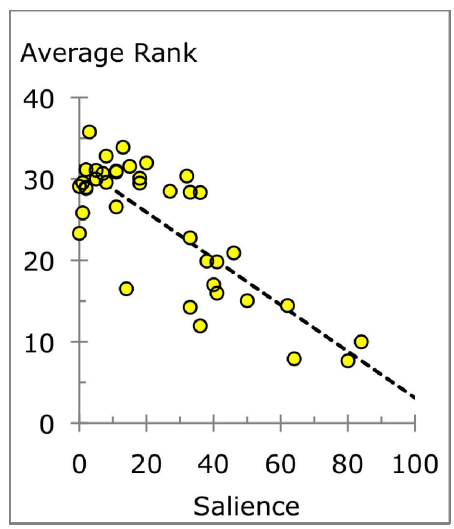

Figure 9. Cross-plot of the metric of salience and the average ranks of the ratings provided by reviewers of video clips containing pedestrian encounters.

DISCUSSION

The primary limitation of the methods is their reliance on the height of a pedestrian as the proxy for risk. This shortchanges children. Accordingly, we plan to revise the metrics by replacing pedestrian height with the distance estimate used by the systems in their detection task.

The two metrics quantify system performance along complementary dimensions. The Metric of Similarity provides a time-trace and composite score of system performance. The Metric of Salience provides a snap-shot prediction of driver acceptance of system output. By applying the metrics, original equipment manufacturers and suppliers have been able to identify factors that contribute to user acceptance of Night Vision systems and their performance.

ACKNOWLEDGEMENTS

This work was supported by a grant from the European Union's 7th Framework Programme under Contract # 216384. Ms Irmgard Heiber was the project officer. Members of the research consortium were Acreo AB, Autoliv Development AB, Daimler AG, Kungliga Tekniska Högskolan, Linköping University, Sensonor Technologies AS, and Umicore SA/NV. The first author worked on the project while a Guest Professor at Linköping University. All opinions in this article are the authors' and have not been officially or informally endorsed by the European Union or by consortium partners.

REFERENCES

Baddeley, S. D., & Hitch, G. 1993. "The recency effect: Implicit learning with explicit retrieval?" Memory & Cognition, 21(2), 146-155.

European Union 7th Framework Programme (2011). "FNIR: Fusing far and near infra-red imaging for pedestrian injury mitigation." Brussels: European Union.

Greene, R. L. 1986. "Sources of recency effects in free recall." Psychological Bulletin, 99(2), 221-228.

Källhammer, J.-E. In press. "Rethinking false alarms by automotive active safety systems."

Källhammer, J.-E., & Smith, K. In press. "Driver Acceptance of Pedestrian Alerts by a Night Vision system."

Källhammer, J.-E., Smith, K., Karlsson, J., & Hollnagel, E. 2007. "Shouldn't the car react as the driver expects?" Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design. Stevenson, WA.

Pavlov, I. P., 1927. "Conditioned Reflexes." Clarendon Press, London.

Pieters, R. G. M. & Bijmolt, T. H. A. 1997. "Consumer memory for television advertising: A field study of duration, serial position, and completion effects." Journal of Consumer Research, 23(4), 362-372.

Seamon, J. G., Marsh, R. L., & Brody, N. 1984. "Critical importance of exposure duration for affective discrimination of stimuli that are not recognized." Journal of Experimental Psychology: Learning, Memory and Cognition, 10/3), 465-469.

Smith, K. 2010. "Quantifying active safety system performance at pedestrian detection." Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomic Society. San Francisco, CA, 2038 - 2042.

Smith, K., Schweiger, R., Ritter, W., & Källhammer, J.-E. 2011. "Development and evaluation of a performance metric for image-based driver assistance systems." 2011 IEEE intelligent vehicles symposium. Baden-Baden, Germany.