Retrieval of Relevant Visual Data in Selected Machine Vision Tasks: Examples of Hardware-based and Software-based Solutions

Andrzej Śluzek

***Abstract*—**To illustrate diversity of methods used to extract relevant (where the concept of *relevanc*e can be differently defined for different applications) visual data, the paper discusses three groups of such methods. They have been selected from a range of alternatives to highlight how hardware and software tools can be complementarily used in order to achieve various functionalities in case of different specifications of "relevant data". First, principles of gated imaging are presented (where relevance is determined by the range). The second methodology is intended for intelligent intrusion detection, while the last one is used for content-based image matching and retrieval. All methods have been developed within projects supervised by the author.

***Keywords*—**Relevant visual data, gated imaging, intrusion detection, image matching.

I. Introduction

MACHINE vision algorithms and methods are already mature enough to be used as substitutes of human vision in many industrial and commercial applications. However, the level of complexity and sophistication in such applications is still limited due to several reasons. The most important reason is apparently a complex mechanism of interactions between *perception*, *seeing*, *recognition* and *understanding* visual contents of observed scenes in a human sense of vision (see examples in [1]). In artificial vision systems (where such mechanisms are at a very basic level, if any) it is still impossible to emulate human vision senses with all neuropsychological functionalities. Because of that, the majority of commercial/industrial machine vision systems are based on a simple multistep scheme consisting of: (1) image acquisition, (2) image pre-processing, (3) image analysis and (4) image categorization (detection, recognition, etc.) operations which may sometimes overlap and/or interact.

The major disadvantage of the above scheme is its limited adaptability and flexibility. In other words, the amount of processed information (especially in the first two steps) depends primarily on the size and complexity of acquired images and not on the nature of a specific problem. The extraction of visual data relevant to the problem is usually performed at later stages, with a large amount of computational power already unnecessarily spent. Therefore, the problem of intelligent data reduction (i.e. retrieval of data that are relevant to a given machine vision task) at the earliest phases of machine vision operations is of fundamental importance, especially for vision systems with limited processing or communication capabilities, and/or working under tight timing constraints.

Numerous approaches and techniques have been proposed for data reduction (relevant data retrieval) at early stages of image acquisition and processing. They range from simple binary thresholding to complex evolutionary algorithms (e.g. neural networks highlighted in [2]) or sophisticated smart cameras (e.g. [3]) trying to "understand" the content of captured images.

As an illustration of diversity in methods used to efficiently isolate the visual data of interest in various problems, this paper discusses three such methods. They have been selected from a range of alternatives to highlight how hardware and software tools can be complementarily used in order to achieve various functionalities in case of different specifications of "relevant data". All methods have been developed within projects supervised by the author (the individual contributions of the team members are highlighted in the acknowledgement section at the end of this paper).

The first method, i.e. gated imaging briefly overviewed in Section II, is used when the relevant data are contained within predefined (but not always known) distances from the image capturing devices. Additionally, the method enhances the visual quality of acquired images so that it is particularly suitable for difficult conditions (e.g. underwater applications). Section III describes results that are intended for visual detection of intruders. A combination of an FPGA-based platform with hardware-implemented algorithms provides a intelligent detector that can locally classify intruders and eventually transmits visual data on "potentially dangerous" intruders. Finally, we show how to adapt a well-known method of keypoint-based image matching for retrieval of images similar to a given query image. The method is individually tuned for each query images so that the most relevant data/characteristics of the query are used for image matching.

A. Śluzek is with Nanyang Technological University, Singapore; he is also on a long-term leave from Nicolaus Copernicus University, Poland (phone: +65-6790-4592; fax: +65-6792-6559; e-mail: assluzek@ntu.edu.sg).

II. Data Reduction by Gated Imaging

Gated imaging (e.g. [4], [5]) is a technique of image capturing by using devices that discriminate in time domain between signals arriving from different sources (e.g. between object-reflected and backscattered light). Typical gated imaging systems consist of a pulse-generating laser (with its pulses usually diverged into a conical shape) a high-speed and high-sensitivity camera with electronically controlled (gated) shutter, and a control and synchronization circuitry. If laser pulses reflected from a object return to the camera with its shutter open, the object's image is captured.

If the reflected pulses return before the shutter is open (i.e. the object is near the camera) or after the shutter is closed (i.e. the object is at a longer distance) the object is not depicted on the acquired image so that it effectively becomes invisible. Thus, by synchronizing the shutter opening timing with the return of pulses reflected from an object, it is possible to capture images depicting only those fragments of the observed scene which are within a pre-selected range of distances. Fundamentals of gated imaging are explained in Fig. 1.

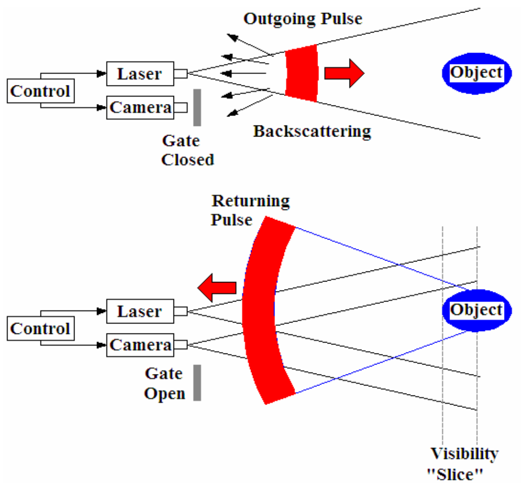

Fig. 1 Principles of operation of gated imaging

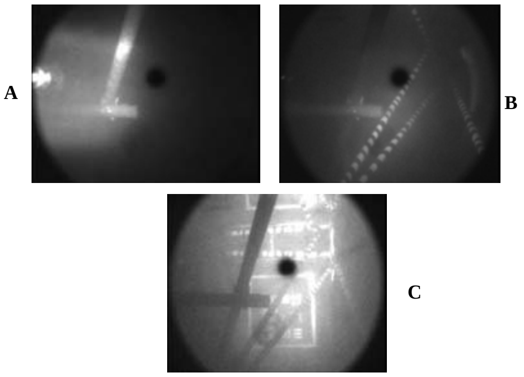

Fig. 2 Gated images of the same scene captured with diversified gate opening delays

Fig. 2 shows three gated images of a laser-illuminated scene captured with a short ((Fig. 2A) medium (Fig. 2B) and longer (Fig. 2C) gate opening delay. The figure is a convincing example of the most fundamental data reduction scheme provided by gated imaging, i.e. by appropriate delays in gate opening and closing, it is possible to capture images containing visual data only from within a pre-selected range of distances.

Typical short-range gated imaging systems have shutter opening periods of 2-20ns duration and correspondingly short pulses generated by the laser illuminator. Thus, the typical depths of "visibility slices" (see Fig. 1) are 0.3-3.0m in air and 0.2-2.2m in water (differences caused by different speed of light in both media). The actual visualization range is jointly determined by the energy of laser pulses and the medium translucency (see below).

A. *Visibility Improvement in Gated Imaging*

Backscattering is unwanted but unavoidable component of all visual signals captured in natural conditions. The backscattered signal reflects uniformly from the whole volume of the medium (e.g. air, water) so that standard imaging devices (e.g. camera with typical shutter speeds) would accumulate a large amount of backscattered noise. Gated imaging device, however, capture only an insignificantly small amount (compared to the amount of signal reflected form the objects of interest) of backscattered noise because of a very short gated opening time. Therefore, gated images not only capture relevant visual data (i.e. data within selected ranges) but also, in high-turbidity media in particular, the quality of captured data is higher. As an illustration, Fig. 3 compares a non-gated image to a gated one captured under the same conditions.

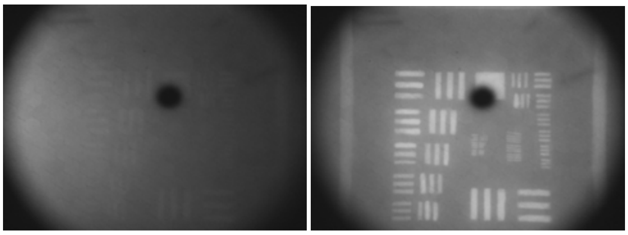

Fig. 3 Comparison between a non-gated image (A) and a gated one (B) captured under identical conditions

In general, gated imaging systems can visually penetrate turbid media 3-6 times deeper than standard imaging devices. However, models of gated imaging (see [5]) indicate that for increased media turbidity shorter illumination pulses (and correspondingly shorter shutter openings) are needed to maintain good image quality. Therefore, the "visibility slices" become narrower the amount of acquired visual data may be unacceptably reduced.

The implemented solution for this problem consists in data fusion from several gated images captured for gradually changing gated opening delays. In order to perform such an operation in real time, a feasibility study FPGA-based controlling & processing device has been developed (see [6] for more details). Exemplary results (presenting individual gated images, the online fusion results, and comparing them to a non-gated image) are given in Fig. 4.

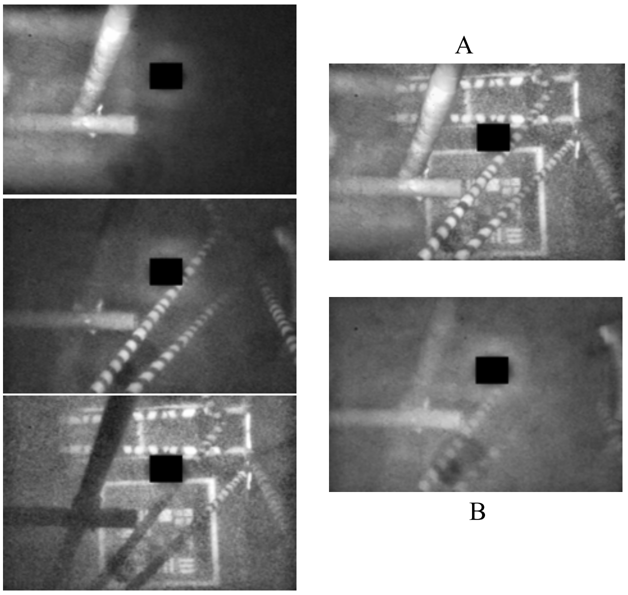

Fig. 4 Three gates images (left column) captured in turbid water; (A) – the image fused from gated images; (B) – the corresponding non-gated

By incorporating such a device into a gated imaging system we can: (1) flexibly define the range and depths of the relevant visual data, (2) capture visual data in turbid media at much longer ranges than standard image-acquisition techniques, and (3) maintain satisfactory quality of the acquired images under a wide range of media turbidities.

III. Extraction of Relevant Data in Intrusion Detection

In monitoring and surveillance systems, the concept of *relevant visual data* is significantly different. Even though the range information may play a certain role, the most important data are image fragments that suddenly change (i.e. a possible indication of intrusions). In case of numerous intrusion detectors interconnected by communication channels of limited bandwidth (e.g. a network of wireless sensor nodes) the amount of such data is usually very large, but we can attempt to minimize it (without compromising the detection reliability!) in particular the amount of transmitted data.

In this section we present a solution that has been developed for semi-automatic security/surveillance systems with a large number of wirelessly interconnected cameras (more details in [7]) where the operator is the ultimate verifier whether and how the intrusion should be handled. To manage the huge information flow and avoid unnecessary data transmission, we incorporate field programmable gate arrays (FPGA) into the camera nodes. FPGA locally processes the visual data and transmits the analyzed information to the destination in case of unusual events (that are preliminary recognized as dangerous intrusion). However, the visual assessment of a situation by a human is still considered the ultimate factor in taking important decisions.

Since the amount of unprocessed data transmitted across the network would quickly saturate the human attention, the extraction of relevant data is the top priority. Image fragments containing suspected intrusions are obtained by background removal. In our method, we use background subtraction with additional image processing algorithms (including mathematical morphology) to compensate for the fluctuations of illumination, minor motions of the background objects (e.g. grass or tree leaves) vibrations of the sensor platform, etc. The background is regularly updated when no intrusions are detected. Additionally, if an allegedly intruding object is not eventually considered dangerous, it will be included into the background. In general, the process of background update varies according to the application, as each application might require a different method of background updates. All methods, nevertheless, are hardware-implemented within the FPGA.

Eventually, after the background removal a region (or several regions) representing the suspected intruder(s) are obtained in two different forms: either as a binary image (intruder's silhouette) or as a masked original image (intruder's image). An example is given in Fig. 5. In general (subject to further criteria highlighted in the following subsection) only intruder's images are wirelessly transmitted to higher levels of the decision system (e.g. to a human operator).

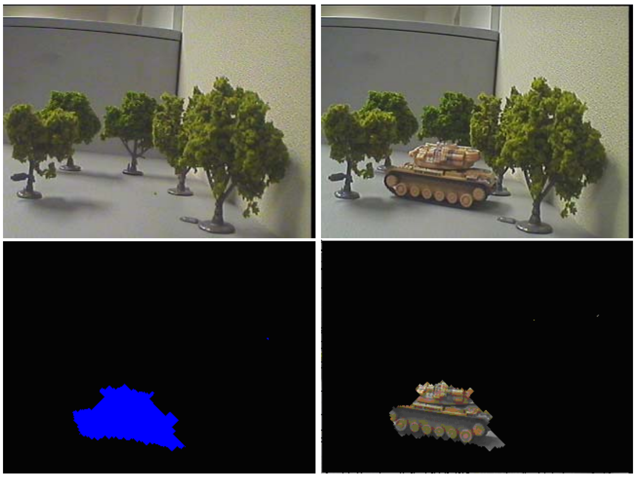

Fig. 5 A background image, an image with an intruder, and the corresponding intruder's silhouette and intruder's image

A. Further Analysis of Detected Intrusions

In selected applications, intruder's silhouettes extracted by background removal can be further characterized by using selected features in order to broadly classify the type of intrusion (actually, silhouettes extracted from a sequence of 2-4 images are used). The proposed features are moments of low order (see [8] and [9] for the general theory of moment invariants) from which several useful descriptors of the intruders can be derived. In particular, the type of intruder's mobility can be broadly categorized by comparing the moments of order 0 and 1 for the silhouettes extracted from two or more subsequently captured images. For example, the relative positions and sizes of two silhouettes shown in Fig. 6 indicate that the intruder is moving to the left and leaving the monitored area.

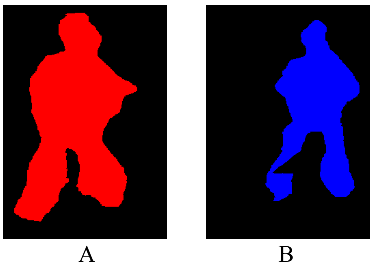

Fig. 6 Silhouette deformations in a sequence of two images

The 2nd and 3rd order moment expressions invariant to resizing, translations and rotations (see [8]) can hypothetically distinguish between living creatures (humans and animals) and man-made objects (e.g. vehicles). The former usually rapidly change their silhouettes when moving, while the latter should maintain approximately the same silhouette for a sequence of quickly captured images. Therefore, significant changes of the silhouettes' moment invariants indicate the intrusion by a leaving creature rather than by a vehicle.

In case of Fig. 6 silhouettes, the overall moment-based conclusion would be: *a leaving creature not approaching the protected area (i.e. a non-critical intrusion).*

The moment-based classification of intrusions is also implemented locally within the FPGA module. Therefore, only images of intruders classified as *critical* need to be transmitted for a visual inspection by a human operator.

The mechanism of relevant-data retrieval proposed in this section is significantly more complex than techniques applied in smart cameras (see [3]). It can be rather perceived as a simple embodiment of a more advanced paradigm of "*seeing only what should be seen*" (which is strongly related to biological principles of seeing and sensing in general, [1]).

IV. Data Relevance in Image Matching

Image matching is currently one of the most important topics in machine vision (in particular in the context of visual information retrieval). At the lowest level, images are typically matched using collections of interest points (keypoints) extracted by various keypoint detectors (e.g. [10]) and subsequently characterized by the corresponding keypoint descriptors (a comparative survey provided in [11]).

Interest points are actually interest regions since they are extracted in a form of (typically) elliptical areas at locations where the image intensities/colours have distinctive local characteristics. The size (scale) of regions indicates how large is the neighbourhood over which these characteristics are the most prominent. Fig. 7 shows examples of interest points extracted by so-called Harris-Affine detector (more details in [10]).

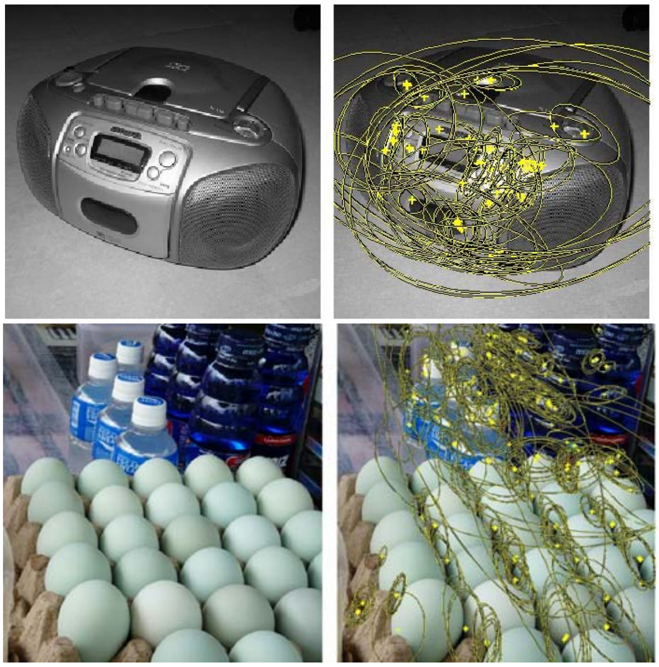

Fig. 7 Interest points detected in exemplary images

Matching the interest point descriptors is a generally used mechanism for determining visual similarity between contents of two (or more) images. However, this mechanism does not take into account the individual properties of matched images. Therefore, we have proposed several modifications that take into account those properties (i.e. we retrieve and exploit the most relevant - in the context of image matching - data).

A. Image Matching by TPS Warping

In this method, we assume that images that depict the same contents (although possibly seen from a different viewpoint and under different illumination conditions) should be related by a certain transformation. The transformation may be a non-linear one (e.g. perspective distortions) but once identified it can be used to bring two images (or at least their common contents) into almost the same form.

The implemented method of image transformation is based on a *thin plate spline* (TSP) warping transformation. TSP transformation (see [12] for more details) is a mechanism for function fitting over a 2D set of scattered points. Assuming a certain set of control pairs (i.e. pairs of points which are known to be their mapping) we build the optimum transformation in a form of an affine transform with non-linear factors added:

$$f(x,y) = a_1 + a_x x + a_y y + \sum_{i-1}^{n} w_i U(\|(x_i, y_i) - (x,y)\|)$$

The control pairs are selected from the interest points which are mutually their closest neighbours (so-called coherent pairs – see [13]). If the images actually depict the same scenes, after the transformations they become very similar. Exemplary effect are given in Fig. 8 (it should be noted that there are always two ways of image warping – image A into image B or another way around – and both are shown in the figure).

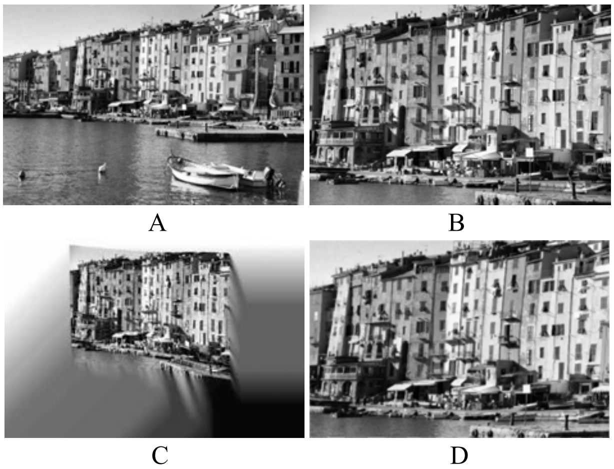

Fig. 8 Two matched images (A and B) and the TPS-warping results. C – Image B warped into image A; D – image A warped into B

The most relevant contents of matched images (i.e. the coherent pairs) enhance performance and simplify implementation of the matching algorithm. In case of actual matches, both images become visually very similar so that matching becomes very simple, while in case of different images a random set of coherent pairs would produce warping that totally distorts the transformed image.

B. *Dedicated Interest Points*

Interest points extracted by standard detectors usually have an unspecified visual semantics. However, in many cases of image matching the presence of similar local geometric structures is the decisive factor in determining similarity between images. Such local structures (e.g. corners, junctions, etc.) are well-proven clues in analyzing the image contents. Unfortunately, they might be often missed by interest point detector. Fig. 9 shows an image where we are interested (for some reasons) in T-junction features. Harris-Affine detector, however, cannot accurately localize any of such features (even though they can be easily identified by a visual inspection).

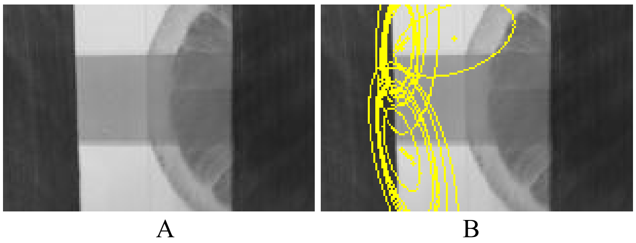

Fig. 9 Exemplary image (A) and its interest points detected by Harris-Affine detector (B)

We propose an alternative method of detecting interest points with visual semantics (i.e. representing locally various geometric features). In the general, the method is based on locally applied Hough transforms (see [14]). Interest points are detected at the locations where the output of the Hough transforms reached a local maximum. Principles of this work are presented in [15], but similar results for junction features only have been also reported in [16].

The T-junction based interest points shown in Fig. 10 are clearly more relevant than standard interest points given in Fig. 9B.

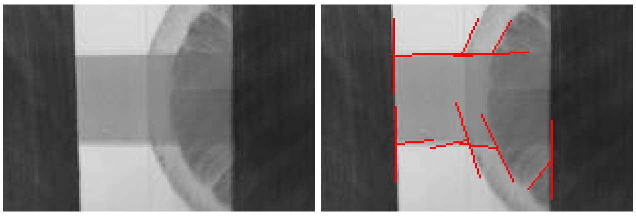

Fig. 10 T-junction based interest point detected in the same image

Results of another exemplary case, shown in Fig. 11, seem also more relevant than standard interest points detected in Fig. 7.

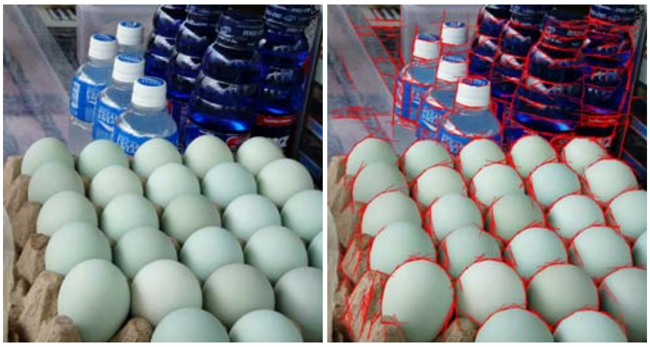

Fig. 11 T-junction based interest point detected in the image shown in Fig. 7

C. *Dedicated Interest Point Descriptors*

Descriptors of interest points are also standardized. The most popular descriptor (and in general the most effective one, according to the study in [11]) is SIFT and its modifications (e.g. PCA-SIFT or GLOH). However, individual images can have some unique characteristics which may be better represented using descriptors individually designed for a given image.

Since a certain level of standardization is unavoidable (e.g. using SIFT as a popular descriptor) we propose a mechanism to build such “individually designed” descriptors over standard ones. Our objective is to have descriptors which discriminate as well as possible between all interest points of a given image by projecting them (e.g. using LDA approach) onto a sub-space with the highest between-class separability and the lowest within-class differences. For such a projection we need a sufficient number of descriptor samples for each class. However, since we consider each interest point of the image a separate class, only one sample is available for all classes. More samples of each class are generated in the following way. First, based on the elliptical region of each detected interest point we produce more elliptical regions with a certain distortion of the ellipse's location, size, orientation and axis ratio. For each such an ellipse, SIFT descriptor is calculated as if it were a regular interest region. Subsequently, the image undergoes several geometric and photometric distortions, and in each distorted image the corresponding elliptical regions are taken into account. Fig. 12 shows an exemplary image with its contrast differently stretched.

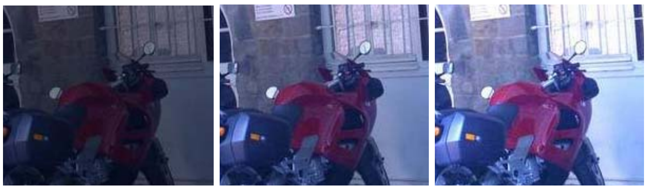

Fig. 12 The image of inerest (centre) distorted by contrast stretching

In this way, a significant number of sample descriptors are created for each original interest point so that the most discriminative (for the given image) sub-space of the SIFT space can be found by using LDA (more details in [17]). Since the applied image deformations are a good estimate of the actual variations expected in other images of the same scene, we apply the same projection of SIFT descriptors for images matched to the current one. Thus, matching is expected to be performed with a higher reliability (the interest points are well discriminated so that fewer false matches are expected) and at lower computational costs (lower dimensionality of the descriptor space). More details of this technique are given in [17].

V. Conclusion

This paper highlighted just a few techniques for relevant data extraction that are effective in selected applications of machine vision. We do not intend to benchmark the presented method or to evaluate their effectiveness in other applications. Our objective is to pass a message that machine vision is still (in spite of over 40 years of development) a developing area with many unsolved challenged and many alternative approaches to those challenges.

Acknowledgment

The author thanks Dr. Tan Ching Seong, Mr. Hikaru Fujishima, Mr. Palaniappan Annamalai and Dr. Yang Duanduan for their vitally important contributions to the results presented in the paper.

We also gratefully acknowledge financial and organizational support from Singapore Technologies Engineering Ltd and from A*STAR Science & Engineering Research Council (Singapore) in our work.

References

[1] A. W. Ellis and A. W. Young, *Human Cognitive Neuropsychology*. Hove: Psychology Press, 1996, ch. 2 & ch. 3.

[2] R. Lepage, "Data reduction in machine vision and remote sensing applications," in *Proc. Int. Conf. Information and Communication Technologies: From Theory to Applications*, Damascus, 2004, pp: LXIII-LXIV.

[3] A.N. Belbachir (ed.), *Smart Cameras*. Springer Verlag, 2009.

[4] G.R. Fournier, D. Bonnier, J.L. Forand and P.W. Pace, "Range gated underwater laser imaging system," *Optical Engineering* , vol. 32, pp. 2185-2190, 1993.

[5] C.S. Tan, A. Sluzek and G. Seet, "Model of gated imaging in turbid medium," *Optical Engineering*, vol. 44, no. 11, 2005, pp. 116002-1-8.

[6] A. Sluzek, H. Fujishima and C.S. Tan, "Real-time digital control in gated imaging," in: *Proc. 5th EURASIP Conf. on Speech and Image Processing, Multimedia Communication & Services*, Smolenice (Slovakia), 2005, pp. 201-206.

[7] A. Sluzek, A. Palaniappan, "Development of a reconfigurable sensor network for intrusion detection," in *Proc.* 2005 *MAPLD Int. Conf. on Military and Aerospace Applications of Programmable Logic Devices*, Washington D.C., 2005

[8] M.K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol.8, 1962, pp 179-187.

[9] S. Maitra, "Moment invariants," *Proc. of IEEE*, vol. 67, no. 4, 1979, pp. 697–699.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. Journal of Computer Vision*, vol. 65, 2005, pp 43-72.

[11] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors." *IEEE Trans. PAMI*, vol. 27, 2005, pp 1615-1630.

[12] F.L. Bookstein, "Principle warps: thin plate splines and the decomposition of deformations." *IEEE Trans. PAMI*, vol. 16, 1989, pp 460-468.

[13] D.D. Yang and A. Sluzek, "Aligned matching: an efficient image matching technique," in: *Proc. IEEE Conf. Image Proc ICIP* 2009, Cairo – to be published.

[14] R.O. Duda and P.E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Comm. ACM*, vol. 15, 1972, pp 11-15.

[15] A. Sluzek, "Images features based on local Hough transforms," *Lecture Notes on AI*, vol. 5712, 2009, pp 143–150.

[16] E.D. Sinzinger, "Amodel-based approach to junction detection using radial energy," *Pattern Recognition*, vol. 41, 2008, 494-505.

[17] D.D. Yang and A. Sluzek, "Performance improvement of SIFT descriptor by using linear discriminant analysis," unpublished.