EG UK Theory and Practice of Computer Graphics (2009)
Wen Tang, John Collomosse (Editors)

# Towards Object Recognition using HDR Video, Stereoscopic Depth Information and SIFT

Michael May[1], Tim Morris[1], Keith Markham[2], William J. Crowther[3] and Martin J. Turner[4]

[1]School of Computer Science, The University of Manchester, UK. [2]MBDA UK Limited.
[3]School of Mechanical, Aerospace and Civil Engineering, The University of Manchester, UK.
[4]Research Computing Services, The University of Manchester, UK.

**Abstract**

*In this paper we propose a framework that will recognise objects from a moving platform using scale invariant features, high dynamic range (HDR) video and stereoscopic depth information. The paper focuses on initial work involving feature extraction from HDR images using SIFT. Initial results show an increase in the number of features extracted from HDR images compared to conventional, low dynamic range (LDR), images.*

Categories and Subject Descriptors (according to ACM CCS): I.4.9 [Image Processing and Computer Vision]: Applications I.4.6 [Image Processing and Computer Vision]: Edge and Feature Detection

## 1. Introduction

Object recognition in computer vision is the detection and classification of target objects within images or video. The human visual system recognises a multitude of objects with little apparent effort, even with large variations in viewpoint, size, scale, rotation and levels of occlusion. This is taken for granted in humans, but it has proved difficult for computers to create context from 2D arrays of pixels.

In this paper we outline a framework for object recognition for use on mobile platforms. Possible practical uses of the system include automatic driving and navigation for road vehicles, mobile recognition solutions or unmanned air vehicles (UAVs). Speed and accuracy are important factors to allow real-time responses to recognised objects. The proposed framework can be broken down into several research objectives:

1. Create a high dynamic range video stream.
2. Locate scale, rotation, illumination and viewpoint invariant keypoints within each image of the video stream.
3. Generate 3D information from the video stream which will allow clustering of keypoints and differentiation between individual objects within a scene.
4. Generate 3D keypoint information for target objects in a database.
5. Match target objects from the database to the video stream using the keypoints and depth information.
6. Track located objects.
7. Recognise changes in a scene from previous images of the same location.
8. Enable communication and collaboration between multiple mobile platforms.
9. Create a system capable of running in real-time.

The framework indicates how the constituent parts will be combined in an attempt to make an effective system for object recognition. For each objective we will attempt to create a new solution or optimise an existing solution. We will initially focus on a subset of the objectives due to resource constraints.

The main focus of this work-in-progress paper is the first stage of research; reliable and robust feature detection from high dynamic range (HDR) [RWPD06] images. Experiments to discover the speed and accuracy of using a feature detector such as the scale invariant feature transform (SIFT) [Low04] to detect features in HDR images is being conducted. The results are then compared to low dynamic range (LDR) implementations. This is important as most aspects of the framework rely on fast and accurate feature detection. An empirical investigation will help uncover the parameters of the SIFT algorithm which produce improved results under different conditions and increase understanding of the best way to generate keypoints for object recognition.



DOI: 10.2312/LocalChapterEvents/TPCG/TPCG09/165-168

**Figure 1:** *The top left image is a tone map of a HDR image created from the three other, LDR, images. Note that in the tone mapped image details from the scene are clearly visible both inside and outside the window.*

## 2. Background on SIFT Feature Detection and HDR

In this section we outline some of the work that is relevant to the primary area of research within the framework; reliable and robust feature detection from HDR images.

### 2.1. Scale Invariant Feature Transform

The original SIFT feature detection algorithm developed and pioneered by David Lowe [Low04] is a four stage process that creates unique and highly descriptive keypoints from an image. Due to the process the keypoints are invariant to rotation and robust to changes in scale, illumination, noise and small changes in viewpoints.

The keypoints can be used to indicate if there is any correspondence between areas within images. Clusters of keypoints from an image that are similar to a cluster of keypoints from another indicate, with a high likelihood, matches between the two areas. This allows object recognition to be implemented by comparing keypoints generated from input images to keypoints generated from images of target objects.

The four stages of the SIFT algorithm are as follows, full details of which are given in Lowe's paper [Low04]:

1. Scale-space extrema detection.
2. Keypoint localisation.
3. Orientation assignment.
4. Creating the keypoint descriptor.

To match keypoints the Euclidian distance between two keypoint vectors is used to find the nearest neighbour. To reduce the time spent searching for the nearest neighbours within a large number of keypoints a Best-Bin-First (BBF) algorithm is used.

### 2.2. High Dynamic Range Images

Dynamic range is the ratio between the brightest and darkest pixels in a scene. A HDR image often consists of three 32-bit floating point numbers [RWPD06]; one to store each channel of a pixel in an image whereas a standard LDR image uses 8-bits per channel. This means a HDR image can store brightness values over a range of 79 orders of magnitude, a range much greater than the 2 orders of magnitude of a conventional image [RWPD06].

In a LDR image data outside the range would be rounded to the nearest value so the difference between two bright objects, one just outside the image's range and one many times brighter, is then lost.

With LDR photography an exposure must be selected to attempt to capture the most important information within the limited dynamic range of the camera. This can be difficult and in most cases information is lost. The large range of a HDR image means that these differences in brightness are almost always captured. In terms of object recognition, it is proposed that more information from the dark and bright areas means that there is a higher probability of locating the object of interest due to the higher number of stable features available.

Although HDR image formats can hold much more information current cameras do not generally capture HDR data. The capacity of CCD sensors in most digital cameras is limited to between 8- and 12-bits [RWPD06] per colour channel which is not enough to capture the dynamic range of a scene.

HDR images are generally generated from multiple LDR images of the same scene taken in quick succession at different exposures as demonstrated by Debevec and Malik [DM97] (Figure 1). The response function of the camera is computed, which maps the pixel value stored in an image to the radiance in a scene. Using this and a weighting function, which reduces the contribution of points at the edges of the dynamic range of the LDR image, a HDR image can be created. The points at the dynamic edges of the LDR range are unreliable as they may not accurately represent the brightness. The HDR image is created from the areas of high detail over all the LDR images.

### 2.3. Tone Mapping

It is impossible to display HDR images on most displays as the dynamic range of the average monitor is only 2 orders of magnitude [RWPD06]. Tone-mapping has been developed to convert a HDR image into a regular 8-bit LDR format so that they can be viewed on a conventional display.

The method invented by Fattal et al. [FLW02] works by reducing the gradient magnitude in the areas of high gradient while preserving the areas of low gradient. The human visual system is not very sensitive to absolute brightness but responds to local contrast meaning that global differences in brightness can be reduced so long as the darker parts of the image remain darker and the brighter parts remain brighter. Reducing the gradient magnitude of the whole image uniformly would remove texture caused by small gradients so to maintain these a weighting is used.

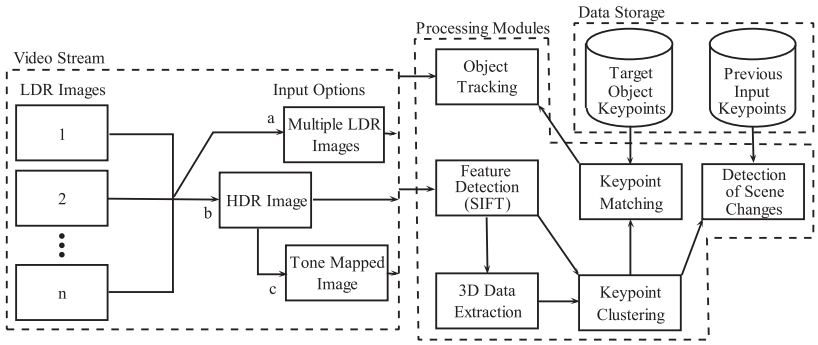


**Figure 2:** *Framework for the proposed 3D HDR Video SIFT object recognition system.*

## 3. Proposed Framework for Mobile Object Recognition

The proposed framework is outlined in Figure 2. The initial input is a series of LDR images. These are used to create one of three options of HDR input; (a) the original LDR images, which can be used to create a set of pseudo HDR keypoints, (b) a full HDR image and (c) a tone mapped image created from the HDR image. It should be noted that the input proposed is in the form of a continuous real-time video stream.

The feature detector will generate keypoints which uniquely describe areas within the image and are invariant to scale, viewpoint, rotation and lighting. Currently SIFT is one of the best solutions [MS05] so for this project a large emphasis will be placed on investigating improving the speed and accuracy of this algorithm. Following this the system uses the keypoint data to generate 3D information from multiple stereo cameras, motion parallax or both. The depth information is important to help cluster the keypoints extracted by highlighting the boundaries between objects. Depth maps of objects will generate more information about their surface and the relative position of keypoints in 3D space. This will allow more discriminative keypoint matching and give an indication of the actual size of objects recognised.

There are two possible matching scenarios that we will focus on in this project. The first involves matching the input to a database of keypoints generated from target images. These keypoints must also have depth information associated with them giving a pseudo 3D model of the object. The second scenario involves matching the input images to previously recorded keypoints from a scene. This will allow differences in the scenes to be highlighted from the last time that the mobile platform was in a specific location. This will be useful in indicating the presence of new undefined objects that the system does not have stored in the database. Once an object has been recognised it can be tracked by the system in consecutive frames of the video stream until it leaves the camera's field of view.

The modularity of the framework allows all the processing to be carried out onboard the mobile platform or different parts may be computed in remote locations from the camera. The system also incorporates the ability to share data between multiple mobile platforms. The database therefore can be held centrally or distributed. Processing at the central location would allow for data to be integrated. For example, if two mobile platforms are near each other superior 3D and keypoint information about the scene could be generated by using images from both.

The framework also allows different areas to be researched and developed individually and new modules can be added if required. As new research is completed and new solutions discovered a module can be easily swapped. Modules can also be used for other projects. The system as a whole will utilise multi-threaded processing techniques for to help achieve real-time performance.

## 4. Current Work

Initial research is being focused on the extraction of features from images and the benefits of HDR over LDR. A HDR image should contain more information thus generating more keypoints. Generating more keypoints does not guarantee better matches but as the extra keypoints will be created from information which is unavailable in the LDR images it is likely to be the case. Objects that would otherwise be hidden in bright or dark areas may be visible. An experiment to test this will use four different inputs and compare the results of SIFT for speed and accuracy. The experiment initially focuses on the number of extracted keypoints. The set of images will be used to create the following inputs:

1. **Single standard LDR images.** A single LDR image with no preprocessing will be used as a benchmark for the comparisons. The central image will be used from the set of images captured using the camera bracketing function. This is an option which allows the rapid capture of multiple pictures at different exposures.
2. **Multiple LDR images of varying exposures taken of the same scene (pseudo-HDR keypoints).** Keypoints are extracted from the LDR images captured using the bracketing function. The keypoints will be matched between the images to align them. In locations where multiple keypoints match a single keypoint with the largest gradient magnitude is kept and the others discarded. The gradient magnitude, from the orientation assignment

**Figure 3:** *The three scenes used in testing (tone mapped) [hdr09].*

stage of the SIFT algorithm, indicates the rate of change in the area around the SIFT keypoint. We expect that the keypoint with the highest rate of change is in the image which is best exposed due to the higher level of detail and clearer edges. Under and over exposed areas of images should have lower gradient magnitudes as the data from the edge of the LDR images have less detail, as explained in section 2.2. This will generate a series of keypoints which we expect will be similar to those generated directly from a HDR image.

3. **HDR images with 32-bit floating point numbers representing each channel.** Each set of output images from the cameras bracketing function is used to create a single HDR image. SIFT is designed to work with 8-bit images [Low04] but we have modified it to work with 32-bit HDR images. The original SIFT parameters are unlikely to be optimal for HDR images due to the increase in magnitude and range of the values. Parameter optimisation is an aim for future investigation.
4. **LDR tone mapped images created from the HDR images.** We are interested in discovering how the tone-mapping process affect the keypoints produced by SIFT. The algorithm is designed to preserve detail so we compare its response with that of a full HDR image. This requires no change in the SIFT algorithm as the tone mapped images are 8-bit.

### 4.1. Initial Results

| | LDR | Pseudo-HDR | HDR | Tone Map |
|---|---|---|---|---|
| Scene 1 | 1767 | 5502 | 7981 | 3508 |
| Scene 2 | 4719 | 10940 | 11944 | 6502 |
| Scene 3 | 1508 | 6997 | 6534 | 3374 |

**Table 1:** *The number of features detected in the 3 scenes (Figure 3).*

Table 1 shows some initial results for feature extraction. For Pseudo-HDR, gradient magnitude is used for keypoint selection. The HDR keypoints are generated using a contrast threshold of 0.007 within the SIFT algorithm. The results show that more keypoints are generated from the pseudo-HDR, HDR and tone mapped images than the standard LDR images.

## 5. Conclusion and Further Work

The initial results are promising and show a large increase in the number of points extracted. We expect that this will translate to more accurate matching especially in scenes with high contrast.

Further work will focus on matching keypoints and recognising objects within scenes. We will take multiple images of a scene with reference pictures taken of objects within it. We will use the SIFT algorithm to extract keypoints and match the objects to the scene. To generate the data we will use a camera with a bracketing function.

The photographs will be taken under an assortment of lighting conditions and with varied viewpoints. The objects will differ in size and texture and will be rotated and occluded by differing degrees. The number of exposures created in the bracketing stage will be varied. This will provide a wide range of data and indicate under which conditions, and to what degree, the various inputs fail to provide a match.

### Acknowledgements

We would like to thank EPRSC and MBDA for funding the project.

### References

[DM97] DEBEVEC P. E., MALIK J.: Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, pp. 369–378.

[FLW02] FATTAL R., LISCHINSKI D., WERMAN M.: Gradient domain high dynamic range compression. *ACM Transactions on Graphics 21*, 3 (2002), 249–256.

[hdr09] Examples of tone mapped HDR images and exposure blending. http://www.hdrsoft.com/examples.html, Apr. 2009.

[Low04] LOWE D. G.: Distinctive image features from Scale-Invariant keypoints. *International Journal of Computer Vision 60*, 2 (2004), 91–110.

[MS05] MIKOLAJCZYK K., SCHMID C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2005), 1615–1630.

[RWPD06] REINHARD E., WARD G., PATTANAIK S., DEBEVEC P.: *High dynamic range imaging*. Elsevier, 2006.