Camera Tracking in Visual Effects
An Industry Perspective of Structure From Motion

Alastair Barber [*]
Double Negative & University of Bath

Darren Cosker [†]
University of Bath

Oliver James [‡]
Double Negative

Ted Waine [‡]
Double Negative

Radhika Patel [‡]
Double Negative

Abstract

The ‘Matchmove’, or camera-tracking process is a crucial task and one of the first to be performed in the visual effects pipeline. An accurate solve for camera movement is imperative and will have an impact on almost every other part of the pipeline downstream. In this work we present a comprehensive analysis of the process at a major visual effects studio, drawing on a large dataset of real shots. We also present guidelines and rules-of-thumb for camera tracking scheduling which are, in what we believe to be an industry first, backed by statistical data drawn from our dataset. We also make available data from our pipeline which shows the amount of time spent on camera tracking and the types of shot that are most common in our work. We hope this will be of interest to the wider computer vision research community and will assist in directing future research.

Keywords: Matchmove, Camera Tracking, Structure from Motion, VFX, Motion Pictures

Concepts: •**Computing methodologies** → **Computational photography;** •**Applied computing** → *Media arts;* •**Social and professional topics** → Automation;

1 Introduction

Camera tracking in visual effects (VFX) is an important task. Errors in this stage can have drastic knock-on effects further down the pipeline. An accurate camera track is essential to being able to convincingly composite Computer Generated (CG) images onto live-action footage, by ensuring that the virtual camera in a render matches the movement (hence, ‘match-move’) of the real camera. The most common way of determining the motion of a live-action camera is through the use of structure from motion algorithms that use feature tracks over multiple frames to calculate the movement of the camera (extrinsic parameters) and internal parameters (camera intrinsics). This is a very active field of computer vision research, with much work being done on accurate 2D feature tracking and 3D scene reconstruction. In spite of this, camera tracking still takes a significant amount of time in the visual effects pipeline, and is a process that requires a large amount of human involvement. Figure 1 shows the percentage of time dedicated to various tasks of VFX production over 6 feature film projects completed over the 2014-2016 time-frame. The duration and time measurements in this work refer to ‘man-hours’, i.e. the actual time taken by a single specialised and experienced visual effects artist to complete the task. These measurements are recorded in a production management system used for costing and scheduling.

The term Matchmove encompasses camera tracking, body and object tracking. Depending on the studio workflow, the time allocated to the matchmove process may include all of these stages. At Double Negative, the times taken for camera, body, and object tracking are estimated and recorded separately, and the times used in this work refer to ’man-hours’ spent exclusively on camera tracking. In this work we are concerned only with camera tracking as the process of determining 3D orientation and movement of a camera using 2D image tracks along with additional information such as set surveys, camera meta-data and on-set notes - and it is this process that we refer to as Matchmoving. The time dedicated to the matchmove process varies significantly from project to project. A project refers to a feature film for which the company has been contracted to provide Visual Effects for. The nature of these will vary greatly depending on genre, the style of a particular director and the complexity of the effects required. Figure 2 shows the different amounts of time given to matchmoving across the feature films from Fig. 1. Also shown in Fig. 2 are details of two other stages, Rotoscoping and Prep. These processes are often grouped with Matchmove to form ‘Roto., Prep. and Matchmove (RPM)’ as a description for the initial stages of VFX production. These stages are always performed early on in the VFX process, as the results obtained from them (Camera Tracks, Rotoscoped Mattes, ‘Cleaned’ Plates with rigging and markers removed) are crucial for almost all other processes to be performed. It is therefore advantageous for these to be completed quickly and also their duration estimated accurately as to enable efficient scheduling and cost control. Having to correct these stages at a later point will have a drastic knock-on effect on the pipeline, as typically the more complex and thus expensive stages such as animation and effects simulation are performed later in the VFX pipeline. Re-doing these later stages because of an error in camera tracking would have a high cost to the production. Figure 2 would also suggest that there is little consistency in the amount of time taken for matchmove compared to other processes, between projects.

VFX work on a particular project is broken down into a sequence of ‘shots’ of continuous camera footage of a particular scene. The length and content of these shots is variable and can range from a few frames ($<$ 1 sec) up to thousands of frames of fast moving footage. Even accounting for the variable lengths of these shots, the amount of time taken to perform a camera track on a shot has a high variance even across a single project. Figure 3 illustrates this across the previous 6 films. Figure 3(b) describes how much shots with differing levels of difficulty contribute to the overall time taken for matchmoving.

Throughout this work we refer to a *camera solve* as the values for

[*] a.e.barber@bath.ac.uk

[†] dpc@cs.bath.ac.uk

[‡] {oj, ted, rap}@dneg.com

DigiPro ’16,, July 23 2016, Anaheim, CA, USA
ISBN: 978-1-4503-4429-6/16/07
DOI: http://dx.doi.org/10.1145/2947688.2947697

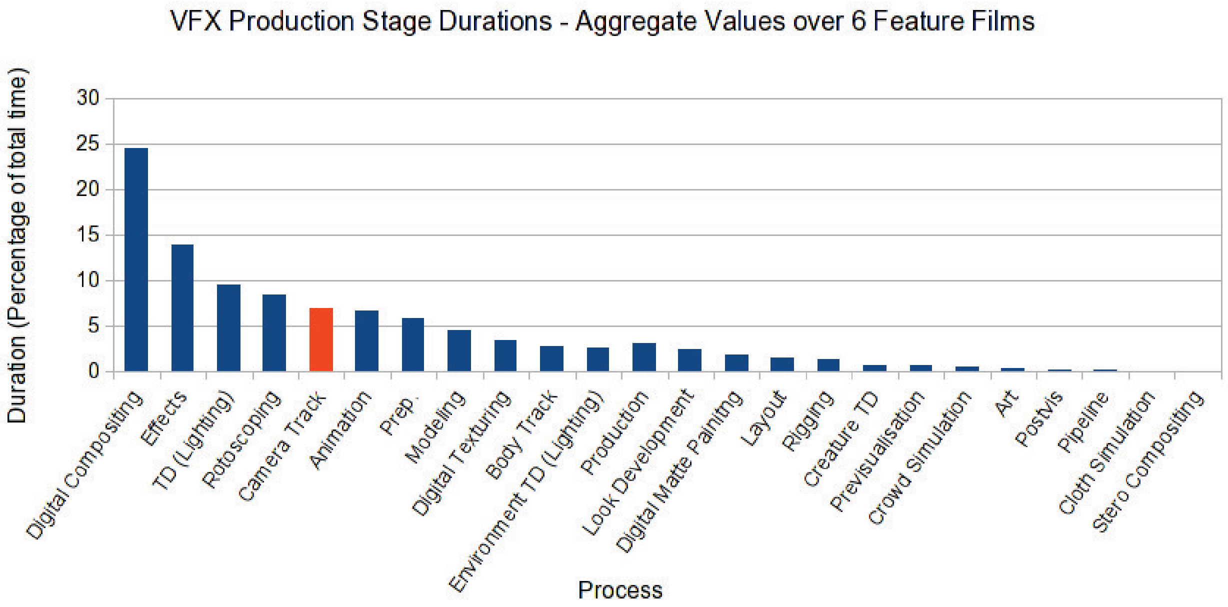


Figure 1: *Duration of various visual effects pipeline processes, with Camera Tracking, the subject of this work, highlighted. This data was taken from an aggregate of the total times taken over the production of 6 feature length films, with the company acting as either sole or a major Visual Effects vendor. For a description of each of these stages see [Goulekas 2010].*

the camera's Rotation and Translation in world space, animated over the sequence. In this paper, we explore the camera tracking process in a major visual effects facility, and the methods used to determine these values. We describe how shots which would be challenging for automated computer vision based methods to calculate a solve for (those consisting of heavy motion blur, wide baselines and low visual texture for example) are handled in production. We also investigate the level of use of automated detection and tracking technologies.

The main contributions of this paper are:

- An up-to-date overview of the matchmove process in a VFX facility and identification of some of the most prevalent shot characteristics which lead to difficulty in providing a solve, and methods used to overcome them.
- Analysis of a large dataset of solved camera tracks from real productions to gauge the impact of these on solve time.
- Suggestions and guidelines that could assist in the scheduling of the matchmove process

The structure of the remainder of this work is as follows: Section 2 describes previous academic literature on the area of structure from motion as well as tools commonly used in industry to implement these methods. In Sec. 3, we discuss the typical matchmove workflow with experienced matchmovers and supervisors. We use their feedback to identify the characteristics of a shot that can cause delays in the process, and how these would be dealt with. Section 4 develops methods for testing for the impact of the characteristics described in Sec. 3 over a dataset of 939 real production shots and shows the results for these.

Assumptions and Expected Limitations The analysis and investigations in this work were conducted exclusively at Double Negative Visual Effects Ltd.- a global visual effects facility dealing solely in Hollywood feature film and high-end television VFX. We recognise that the processes and techniques may be different in organisations dealing with other types of VFX work such as commercials. It is also assumed that all of the shots that are analysed were completed by an experienced matchmover, with levels of skills suited to the type of shot assigned to them. None of the shots analysed in this work were shot in stereo. We expect that some of the information presented in this paper will be well-known to experienced matchmove artists. The objective of this work is to disseminate this knowledge to the wider VFX and Computer Vision research community and also to explore quantitative methods for supporting anecdotal evidence reported by experienced matchmovers. All of the quantitative data reported was gathered from live production data. We acknowledge that a significant limitation of this approach is that we are unable to control for all factors which might make a shot difficult to solve. We discuss examples and likely impacts of this in our conclusions.

2 Related Work - Structure from Motion & Matchmoving

A great deal of work has been performed in the area of Structure from Motion (SfM), and this is a very active area of research. The goal of this procedure is to use multiple frames of 2D images captured by a moving camera to determine 3D scene structure and calculate the trajectory of the camera over time. There are several ways in which this can be achieved, many of which are implemented into commercial software and are used daily by matchmove artists on all manner of productions. One common method for determining a camera's motion is the use of *known correspondences* between images. In this method, a set of image features is tracked over time and their trajectories used to calculate their 3D positions and camera position. Given a sufficient number of correct feature correspondences across two frames, it is possible to calculate the translation and rotation of the camera between frames. An in-depth description and analysis of methods for performing this calculation is available in [Hartley and Zisserman 2004]. In all cases of determining structure from motion, feature points need to be selected and then tracked from frame to frame. Manual feature selection and tracking is a tedious and time consuming task, and there has been much work in developing methods for automating this. One popular feature descriptor and detector is the *Scale Invariant Feature Transform* (SIFT) [Lowe 2004]. This method is particularly robust to large changes in image scale, and can also be used to reliably estimate stable feature points for tracking. It is one of the most widely used methods for describing and matching features selected for tracking, although several other methods exist [Bay et al. 2008][Rosten et al. 2010][Dalal and Triggs 2005]. An earlier work by [Shi and Tomasi 1994] proposes a method for which features can automatically be selected to maximise the number of correct tracks. One of the biggest problems for all feature detection and matching algorithms is their level of robustness under challenging conditions such as motion blur, point occlusion, wide baselines, poor illumination and low visual texture. Many recent works have attempted to combat this. In [Jin et al. 2005], the authors present a method for tracking features in the presence of motion blur in a computationally inexpensive way. The authors of [Huang and Essa 2005] propose a method for tracking through occlusions by segmenting the image into regions representing objects in a scene and building a model for their spatial distribution in the scene from frame to frame. The 'DAISY' descriptor, from [Tola et al. 2010] uses an approach similar to the SIFT descriptor to produce a dense feature track across wide-baseline scenes. Dense feature correspondence, often referred to as optical flow, is the process whereby every pixel in a frame is mapped to a corresponding location in the next frame by an $[x, y]$ motion vector. Motion vectors from optical flow are often used in various tasks in Visual Effects. For example, sequences can be 'retimed' using motion vectors to synthesize additional frames in between existing frames, giving the appearance of slowing down the footage whilst maintaining the frame-rate. The authors of [Xue et al. 2015] use motion vectors to detect and remove reflections from footage. The iterative nature of determining optical flow and additional constraints it places on pixel motion vectors, described in the original paper by [Horn and Schunck 1981] make it a robust method for determining motion in a scene. Later works by [Zhang et al. 2012] and [Tu et al. 2015] propose robust and stable methods for determining optical flow in the presence of heavy motion blur. Whilst robust, calculating dense point correspondence is expensive and unsuitable to perform on many high resolution frames. Furthermore, it would still be necessary to distinguish between movement of objects in a scene and movement brought about by camera movement, for which there would still be a need to identify features in a scene that are static. Methods that make use of additional camera mounted hardware, such as inertial measurement units (IMUs), are also proposed as a method for determining scene structure and camera motion in cases where reliable feature tracks are difficult to achieve. In [Klein and Drummond 2004], a method of using inertial sensor data to adapt the feature description and detection based on an estimation of motion of the camera obtained from the IMU is proposed. In [Okatani and Deguchi 2002], an orientation sensor is used to reduce the amount of point correspondences needed to determine an estimate for camera motion to 2 feature matches, as opposed to at least 6.

Another method for determining camera motion is using 'known shape', most commonly used for camera calibration (e.g. in [Lowe 1991] and more recently [Zhang 2000]). This makes use of real-world measurements taken of a scene to assist in calculating the camera's motion. These can be gathered using survey stations, a LiDAR scanner or photogrammetry from reference cameras. In order to calculate camera motion using this information 2D image points must be registered to these 3D measurements. This will often be done manually on keyframes and camera movement curves interpolated (should the intrinsic parameters of the camera, such as focal length, pixel size etc. be known) and can give very accurate estimates for camera movement in the case of feature occlusion or heavy motion blur. The advantage of this method is that the software need only calculate camera position in 3D space as opposed to calculating 3D locations for a set of point tracks along with camera position. This will also mean that fewer 2D tracks are necessary to solve the camera motion. A disadvantage of this approach is that the calculation for camera movement relies heavily on the values for 3D position being exactly correct and is therefore very susceptible to noise. Furthermore, if the 2D image plane is not exactly aligned to the 3D model the camera solution will not match the footage exactly. 3D information of a scene can be useful in constraining candidate feature tracks, and for determining scale in a scene. [Dobbert 2013, p.211-213] describes how 3D measurements taken on set can be used to aid the process of obtaining an accurate camera solve, and we explore how this is used in the remainder of this work.

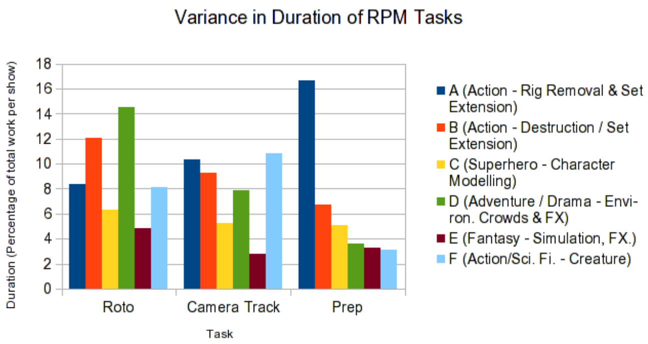
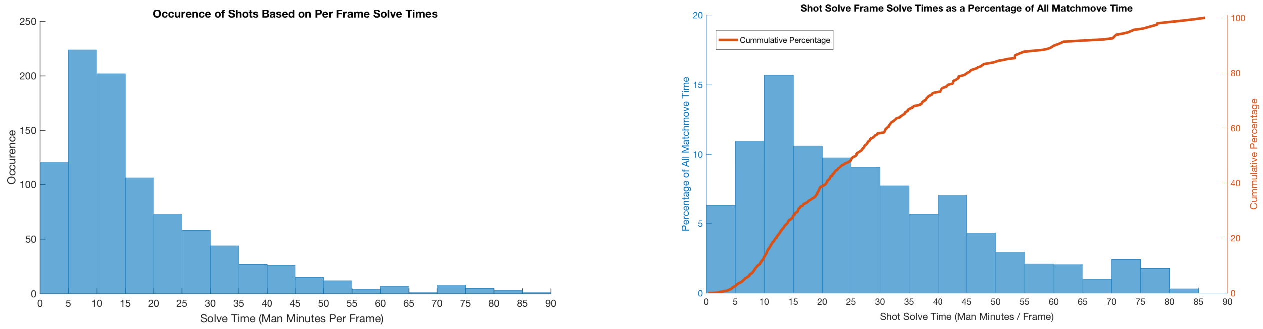


Figure 2: *Chart showing variance in the duration of the Rotoscope, Prep and Matchmove (RPM) Tasks across different film projects of a similar nature - with the type of film and main tasks listed.*

Feature detection and matching within VFX remains an active research area. The authors of [Bregler et al. 2009] specify that 'outlier situations', that would cause many of the previously mentioned automatic feature detection and tracking algorithms to fail, are common in visual effects work. In the software described by [Bregler et al. 2009], a user selects a region to track at key frames in order to train the algorithm, with the software estimating in between morphs for these points in a method based on the Active Appearance Models (AAM) approach (see [Baker and Matthews 2004] for a detailed description of the AAM approach). Conversely, the software package Boujou [Vicon] based on the works by [Fitzgibbon and Zisserman 1998] and [Torr et al. 1998], aims to solve camera motion entirely automatically for any shot. In the following section, we investigate the extent to which automatic solutions are used compared to user-aided camera tracking processes.

3 Matchmove Techniques and Expert Feedback

Despite the use of advanced feature detection and matching algorithms in common matchmove packages, matchmoving in a visual effects facility is still an inherently manual process, as indicated by the amount of man-hours spent tracking a shot shown in Fig. 2. Typically, a matchmover will manually choose 2D feature points in a scene that would appear across consecutive frames and remain at a constant location in world space. A package such as such as 3DEqualizer[Science D. Visions], Autodesk Matchmover [Autodesk] or Boujou [Vicon] would be used to to track these features and use an algorithm such as those detailed in Sec. 2 to compute 3D camera



(a)

(b)

Figure 3: *Charts showing the occurrence of and amount of time spent tracking shots. Chart 3(a) shows that most of the shots in the dataset took between 5 and 10 minutes per frame to solve, followed by those taking 10 - 15 minutes per frame. After accounting for the shot lengths, the highest proportion of time was spent solving shots taking 10 - 15 minutes per frame to solve, followed by those taking 5 - 10 minutes, as shown in chart 3(b) The cumulative percentage line in chart 3(b) shows that despite their lower levels of occurrence shown in chart 3(a), shots taking greater than* 25 *minutes per frame to solve account for approximately* 50% *of all solve time.*

motion from these. Our discussions with matchmove supervisors indicate that automatic feature detection functionality of the software based on computer vision methods is rarely used. One of the main reasons cited for this was that these automatic methods tend to produce a high number of erroneous feature tracks, leading to inaccurate solutions for a 3D camera track. Correcting these would be a time consuming process and it is regarded as more efficient to simply use the time to manually select good points.

Experienced matchmove artists, across both film and television departments at Double Negative were surveyed as to what factors in their experience would cause a delay in creating an acceptable 3D camera track. They were also asked about the stages of a typical matchmove work-flow for a shot. These responses are detailed in the remainder of this section.

3.1 A Typical Matchmove Work-flow

Material Ingestion The first stage of the matchmove process (and all VFX work, after the bidding and awarding processes have concluded) will be when the visual effects facility receives turnover from the studio of live action footage from set. This will almost always be accompanied by a 'lineup sheet' which gives information on the number of frames, frame rate and a short VFX briefing. Of most relevance to the matchmove department will be a description of the camera move and information regarding the lens used. Depending on the production, additional on-set data may have been captured and made available to the VFX facility. This can include:

- 'Witness Camera' footage of the set and main camera during filming
- Lens Grids - Images of a checkerboard calibration target taken with the same lens and camera type as that used in the shot
- Still reference photography of the filming location
- Drawings and measurements taken of the set
- LiDAR Scans of the set and props
- Camera Metadata - such as the focal length, focal distance and shutter angle, synchronised to the footage.

The availability of this data is by no means guaranteed and can sometimes be impossible to obtain (especially in the case of underwater or aerial shots). Recent advances in hardware development have however lead to more data being available more frequently. On large scale productions it is common for LiDAR scans to be gathered due to the relative speed at which they can be acquired and their use in other parts of the production pipeline. Advances in camera technology (for example the Arri LDS System [Arri]), and auxiliary on set data recorders mean that camera data is often also recorded - although care must be taken to ensure that this is correctly synchronised and made available to matchmove artists.

Scheduling It is the responsibility of a matchmove supervisor to estimate the amount of time that will be needed for an artist to track a shot. At this point the contents of the footage, along with the VFX briefing and additional on-set data, is taken into consideration. Often a template specific to a particular facility is used to create a consistent estimation and bidding model. It is at this stage that the 'difficulty' of a shot can be thought to be assessed, with harder shots taking longer for an artist to solve. Discussions with supervisors and artists have identified the following factors that are commonly found to affect the solve time of a shot:

- *Constraints on Camera Movement* A camera that is locked off or moving along or around a single axis will typically be easier to solve than one allowed free movement, or in the worst case handheld or vehicle mounted with high frequency motion.
- *Parallax* The lack of significant parallax in a scene can cause difficulty in solving a scene. It should be noted that the impact of this is often considered in relation to the camera move and its known constraints. For example, a pure nodal movement will not exhibit parallax in a scene. The most difficult types of shot with regards to the level of parallax would be those with a camera movement around an off-nodal point close to the nodal point of the camera. In this case the shot cannot be solved as a pure nodal rotation and parallax is needed to determine the translation of the camera. If this translation is small - the amount of parallax in the scene will be low and hence accurately solving camera translation is challenging.
- *Focus Pulls and Lens Type* Changing focal distance will often result in tracked features in parts of the image becoming blurred and difficult to track. Changes in focal distance will cause a change in magnification of the lens. Specialist lenses, such as anamorphic lenses, introduce complex lens distortions which vary with changing focus and zoom. These changes are not symmetrical across both horizontal and vertical axes, and causes an effect sometimes referred to as 'anamorphic breathing'. This asymmetry can also introduce other distortions into the image and is difficult to accurately model. For this reason, along with other distortion introduced by anamorphic lenses, footage filmed with anamorphic lenses will often be expected to take a longer

time to solve than the same shot filmed with a spherical lens would. Modern lenses are developed to have smaller amounts of distortion. However lens distortion introduced by older lenses is sometimes regarded as artistically satisfying and this is what the filmmaker will base their decision on when selecting a lens for a shot.

- *Availability of 2D Features* The amount of trackable 2D features will depend on factors such as illumination, the scene being filmed and any foreground objects or movement causing occlusion. If features are well defined and present throughout the shot automatic tracking software can be used effectively to track features accurately and give a good solution for a 3D camera motion based on 2D feature tracks.
- *Motion Blur* is common on many shots and is a function of the camera movement. Large amounts of motion blur can add days on to the amount of time needed to solve a shot due to the fact that blurred 2D features are much harder to track automatically.
- *Availability of On Set Measurements and Meta Data* Although the amount of data captured on set in the forms of survey measurements, LiDAR scans, and camera metadata has increased, there is still a great variance in it's availability from shot to shot. Filming conditions, financial and time constraints mean that the availability of this data is never guaranteed. Furthermore - it is often necessary to perform processing on this data prior to its use in the matchmove work-flow, which can be time consuming and also prone to the introduction of error. Usually lens grids are shot as a standard procedure for films using visual effects. However they may sometimes be unavailable or shot incorrectly in which case an approximation of the lens distortion must be estimated – adding time to the process.

One of the most important pieces of information gathered from discussions with matchmovers was that each of these points are rarely taken in isolation, and that the general process of matchmoving is often seen as a problem solving exercise as opposed to a pure computer vision and structure from motion problem. It is rare that a shot will be solved using 2D feature tracks alone. Knowledge of the scene, reference footage and any available meta data will all be used if available and combined to produce a camera solve.

2D Tracking One of the first stages in solving a camera movement is 2D tracking, whereby an artist selects important features in a start frame and tracks these over the footage or until they disappear from view. Many works in the computer vision domain propose methods for automatically detecting good candidate features to track (see section 2). From our discussions with matchmove professionals there was unanimous conclusion that these are almost never used in the matchmove process, despite their inclusion in popular matchmove packages. Matchmovers found that they usually produce poor candidate track points and there is little time saving in having to check and correct automatically detected features tracks as opposed to manually selecting features to be tracked. An experienced matchmover will be able to determine parallax in a scene and select points at different depths in order to take advantage of this. This is something that automatic methods will be unable to do although it may be possible for these methods to determine this based on their 3D solve of the scene. Automatic feature detection will also struggle to differentiate between items moving in the scene and static features which is again something that an experienced matchmove artist can do quickly and accurately.

The types of features selected for tracking will vary between shots and conditions of the plate, and the software package being used. Often corners and prominent details are selected. However, it was reported that with modern software packages good results can be obtained from tracking large surfaces as pattern areas, and these are often tracked more consistently and accurately throughout a shot. The presence of heavy motion blur in an image can cause difficulty in obtaining accurate camera tracks. Motion blur is commonly present in footage as the shutter opening period in motion pictures is traditionally almost always half of the frame duration. It was often felt that with the availability of on-set survey data motion blur is easily overcome by using the 2D to 3D registration technique (motion from *known shape*) described previously in Sec. 2. In this case, lining up the 2D image at keyframes to the 3D model and solving 2D tracks as a minimisation problem can produce good results. However, in the absence of this information it is acknowledged that significant motion blur, particularly in the case of hand-held cameras can contribute to a significant increase in the amount of time taken to solve a shot.

Modern matchmove software will also allow a user to input approximate 3D locations or 2D distances in real world units between points and indicate that they lie on a shared plane, which can be used by the 2D tracking algorithms to constrain the search for feature matches. For this to be most effective accurate lens distortion and intrinsic camera parameters, including the focal length at that frame, must be known. Ideally these parameters will be calculated from lens grid footage using standard camera calibration algorithms. However, in the worst-case, when these aren't available, artists can use known or approximately known shapes in the footage to determine approximate lens parameters.

3D Solve and Lineup It is increasingly common for camera movement and position to be represented in real-world absolute coordinates and units relative to a set. Doing so allows for other departments in the visual effects pipeline to work consistently, i.e. a virtual 3D model can be drawn and scaled to an accurate size relative to the real set. For this to happen it is necessary to take a survey of the set. In the best case a tool such as LiDAR, or laser surveying systems, can be used to generate a highly accurate and dense point cloud of objects in the scene and this 3D representation can be aligned with tracked 2D image points. Even if detailed scans and measurements are not available artists will often search for clues as to the dimensions of objects or locations in a scene online or from set and prop plans, if available. Most matchmove packages will allow for simple 3D geometry to be either drawn or imported into the scene and positioned on the 2D image plane for the artist to use as a reference for alignment and checking the accuracy of a proposed 2D track. [Hornung 2010, p. 45-68] gives a good overview of a typical process of using online resources and measurements to estimate the location and displacement of a camera in a scene in real-world values. While LiDAR scans are often considered the most useful form of 3D scene geometry available it was reported by more experienced matchmovers that care must be taken in its use. LiDAR scans are often acquired at a very high resolution which can be difficult to work with. They therefore must be down-sampled (*decimated*) and processed for use in a studio's production pipeline, which can introduce error. LiDAR scans also exhibit occlusion, whereby objects closest to the scanner will cast a 'shadow' across the scene. LiDAR scans cannot also locate 2D tracking markers placed on set. These high-contrast features will be placed to assist 2D tracking and will be removed in the plate cleanup stage. When used correctly they can be very easily tracked using computer vision methods. It was felt that, ideally, a matchmove supervisor from the VFX facility would be on set during filming and have the ability to measure props, scenery, actors and the location of tracking markers along with a LiDAR scan. Unfortunately, with filming schedules being strictly controlled it is not always practical for this to happen. Advances in technology mean that LiDAR scans are being obtained more regularly, at a lower cost and can be performed fast enough to fit in with filming schedules.

In its simplest sense, the process of lining up a camera‘s tracked 2D points to known 3D locations is often approached as a minimization problem. This formulates a solution for the camera movement that will align 2D image points with 3D points in a scene by minimising the error between 3D locations when projected by a candidate camera solve to 2D points and the tracked 2D feature points in the scene. This process can also be used iteratively to assist with establishing a good 2D point track in the case of motion blur or occlusion or to estimate parameters that were not available originally - for example lens distortion. If the geometry of objects in the scene is known, and in particular perpendicular or parallel edges, then the distortion of a lens can be estimated for a single frame.

Solve Assessment Before a camera track is *published* for use in the next stages of the VFX pipeline it must first be approved, usually by the matchmove supervisor for the particular show or sequence. This would be performed manually and visually - most commonly through the use of a 'cone render', where 3D cones are placed in the frame at 3D points around the image (Fig. 4). This render is then played back and if the cones appear as if they were part of the scene convincingly the matchmove can be approved. An error in the camera track will cause the rendered cones to drift or 'float' over the image. Errors in the positions of the cones at this stage would very likely be representative of a poor 2D track or errors in the camera solve so it is expected that cones will (appear to) remain completely static in the scene prior to the track being approved. Additionally the camera tracked footage will be lined up with the 3D scene drawing (a 'wireframe') of the set and this will be checked to ensure that objects in the footage align accurately with the model. This process will take into account the nature of the shot and the required VFX brief, so need not always be absolutely perfect if convincing integration of CGI and photography allows.

A measure of the distance between 2D image points and 2D reprojections of solved 3D points, commonly referred to as the deviation, can be also be used to assess the quality of a 2D track and 3D calculation. However, all matchmovers interviewed agreed that whilst a useful guide to determining the accuracy of a track, taken in isolation it is not a sufficient measure of the accuracy of the overall camera matchmove.

3.2 Summary

From conversations with matchmove professionals, it has become clear that it is not correct to think of matchmoving as purely a computer vision problem. It also became apparent that the available fully automatic solutions are hardly used. One of the most important pieces of information used by a matchmover that can be gathered from a set is accurate measurements of the scene as not only can they be used to ensure that the camera movement is scaled correctly but they can also be valuable tool in helping to determine 2D feature tracks when computer vision based methods fail. In the following section an investigation into a large data set of solved camera tracks is performed in order to assess the impact of having differing levels and types of on set survey data.

4 Quantitative Investigation

As previously mentioned in Sec. 3.1 the amount of time allocated to solve a shot will increase with the estimated difficulty of the camera track. It has also been noted that, given appropriate information about a set, shots which would typically be considered difficult to track from a computer vision point of view can be solved with greater ease than those without this additional information. In this section we perform an analysis on approved and published shots over six shows with varying lens types and amounts of on-set data available with each containing a variety of different camera moves. In total 939 shots were analysed, across 6 feature film show projects for which Double Negative were a lead or major vendor, in the 2014-2016 timeframe.

For each shot to be assessed the approved animated 3D camera position was obtained from the production asset database along with the lens distortion and camera intrinsic parameters. Also gathered from the production management database was the time logged taken to track each shot along with a list of assets available, such as LiDAR scans, prop scans or on set measurements. The time taken to solve the shot, which is being used in this work as an indication of difficulty, is normalised to minutes per frame for each shot to enable fair comparison between shots of differing lengths. The 2D image size for each shot was also scaled to a common resolution (2048 pixels across the longest edge, maintaining image aspect ratio) for each shot in order to take into account the differing formats between projects. The attributes we intend to assess the impact of in this work are:

- Camera Speed
- 3D Scene Data Availability
- Lens Type: Anamorphic or Spherical
- Camera Motion Constraints

As the data is taken from real production footage it is not possible for each characteristic to be tested in isolation in terms of its impact on the solve time per frame. There will not be two shots that would be identical but with a single different camera attribute. For example, although considered more difficult to solve for due to the distortion introduced by the lens, a shot from an anamorphic camera moving at slow speed through a scene with full LiDAR 3D scans would likely take less time to solve than a spherical, hand held shot in a scene with little 3D information available. In order to compensate for this, we test each factor by analysing its effect on the relationship between point velocity and solve time per frame. We also group shots by the level of scene data available for each test. For all the attributes we intend to measure, camera velocity expressed as 2D point velocity (described below) is plotted against solve time per frame and a first-order fit is calculated. The gradient of this fit plotted as a line would be an indication of how the factors would contribute to the solving process, with a shallower gradient representing an 'easier' solve. However, as shot solve difficulty can be influenced by a number of factors, we expect the fitted lines to indicate general trends as opposed to definitive causal relationships.

4.1 Representing Camera Speed as Point Velocity

From discussions with matchmovers and knowledge of computer vision tracking methods we reason that faster camera movements will take longer to solve than slower movements. Assuming a constant shutter speed, motion blur will be more significant in faster camera movements, and can make accurately tracking features difficult. Features will also be visible for fewer consecutive frames, meaning more time will be needed to select features for tracking - resulting in a less accurate estimation for camera movement.

In order to determine a general measure for the velocity of a camera for a particular shot we calculate the projected 2D image coordinates of the 3D virtual cone assets used to assess the shot (see Sec. 3.1 - *Solve Assessment*). Due to the method of asset publishing, and the fact that 2D tracking work is completed across numerous sites worldwide it was not possible to obtain the original 2D track points used to calculate the 3D camera movement. The mean absolute velocity for these points over all frames is then calculated. This *2D Point Velocity* is then used as a measure of camera velocity in this work for comparing camera speeds. Figure 5 Shows the results for shot difficulty as a function of point velocity. It can be seen that in general an increase in the 2D point velocity does lead to an increase in shot solve time.

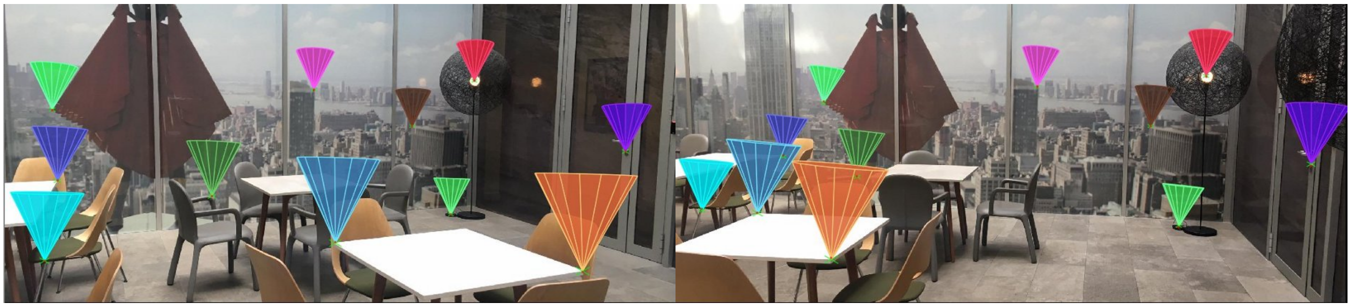


Figure 4: *'Cone Render' for Evaluating a Camera Track.*
Shown are two frames from the same sequence. For the 2D track to be considered successful - the cones should appear as 3D objects in this scene and remain in the same scene location in each frame.

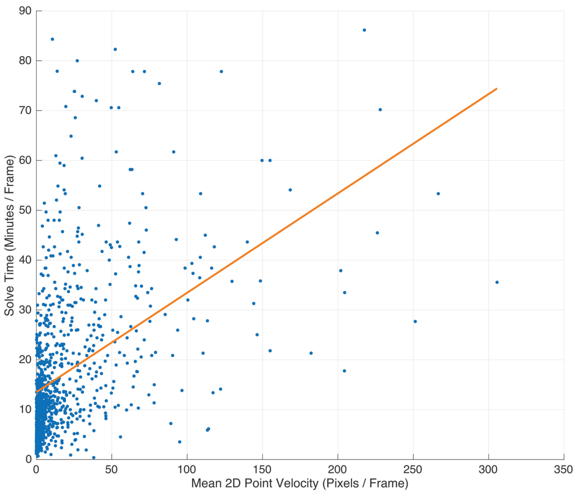


Figure 5: *Matchmove Solve Time and Mean Point Velocity. If increasing point velocity were to lead to an increase in shot solve time, the data points would be expected to tend to a curve with a positive slope.*

4.2 The Effect of 3D Survey Data Availability on Solve Time

Much discussion has been given to the use of 3D on set measurements, in the form of a LiDAR scans or simpler measurements of objects in the scene, as a method for helping solve challenging camera movements. In order to investigate this we group our shot dataset into three groups. Those without any 3D on set data measured, those with simple measurements used to create 3D proxy geometry and those with full dense LiDAR scans available. The method in Sec. 4.1 is then repeated for each group and the fitting algorithm is applied to each dataset in order to determine the trend between point velocity and solve time.

	Lens Type		
3D Scene Data	All	Anamorphic	Spherical
No Survey	16.66		
Proxy Geo.	20.05	17.26	27.51
LiDAR Scan	27.08	11.29	39.77

Table 1: 75^{th} *Percentile* $2D$ *Point Velocities (pixels/frame) for each combination of lens and level of survey data.*

Figure. 6 shows solve time against average 2D point velocity for shots with differing levels of 3D scene geometry data being available. These charts show that for all camera speeds encountered, having a greater amount of 3D survey data of the scene will lower the per-frame solve time of a shot. However, it would appear that there is a tendency for shots with larger amounts of 3D scene data to have a greater point velocity, as is shown by the increase in the 75^{th} percentile values and variance of 2D point velocity with the amount of scene data available (Table. 1).

4.3 Anamorphic and Spherical Lenses

As mentioned previously a 'lineup sheet' usually accompanies scanned footage with information about the camera formats used. Notes from on set, such as a camera data sheet will also usually include information on the lens used. As the choice of anamorphic or spherical lens will have a fundamental effect on the look of the film and would be a decision taken by the filmmaker early on in the process, this information is usually well communicated. In any case the aspect ratio of scanned footage will clearly indicate the lens type. This information is stored as meta-data in the published camera-solve and can therefore be reliably retrieved in an automated fashion across our dataset. Each shot in Fig. 6 is colour coded as shot with an anamorphic or spherical lens, and it can be seen that there are very few anamorphic shots without any 3D scene data, and these are all at a low velocity.

Figure 7 Shows the relationship of point velocity to solve time, for scenes with proxy geometry and LiDAR scans available, with samples split into those shot on an anamorphic lens and those from a spherical lens. There is an insufficient number of shots at different velocities in the dataset with no 3D scene data to reliably estimate separate correlations between point velocity and solve time for both anamorphic and spherical lenses. For each level of 3D data availability the results of Fig. 7 would suggest that, in general, anamorphic lenses do increase the amount of time taken to solve a shot. It can also be seen in Table. 1 that anamorphic shots have a lower general velocity than spherical shots, indicated by their lower value for the 75^{th} percentile velocity. By their nature, anamorphic lenses are physically large and heavy, high cost, and usually add much more barrel distortion to the image and will therefore less likely to be used in shots which require a fast camera movement (especially in the case of handheld cameras). The difference in gradients of the line of fit for anamorphic and spherical lenses is smaller for shots with LiDAR data, which would suggest that having a LiDAR scan could mitigate the impact of anamorphic lenses on solve time. The range of values for solve times for similar 2D point velocities in Fig. 7 is large, even after considering lens type and 3D scene data available. This would suggest that there are further factors influencing the difficulty of a shot.

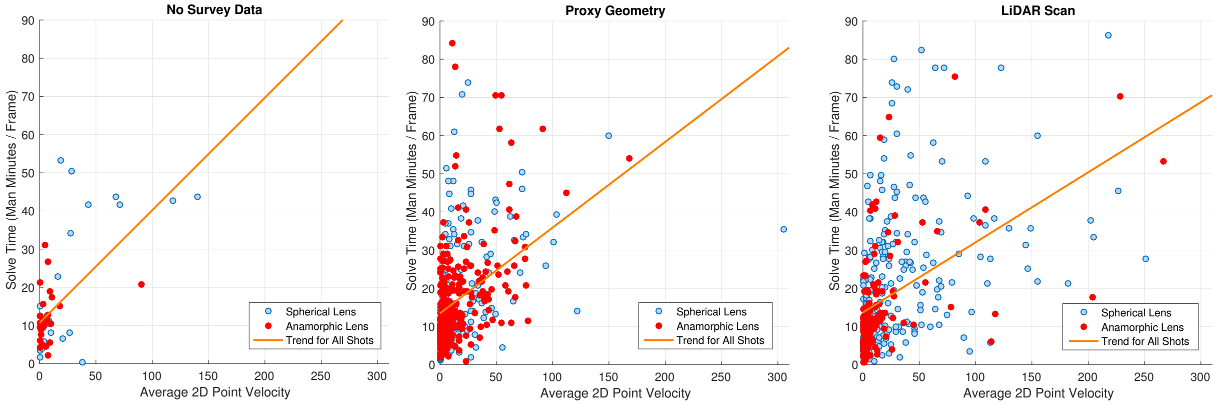


Figure 6: *Solve Time vs. Point Velocity for Different Levels of 3D Scene Data Availability and different lenses. If the availability of more scene data decreases the amount of time taken to solve a shot, the gradient of the orange trend line for LIDAR would be shallower than that of Proxy Geometry, which in turn should be shallower than that for no Survey Data.*

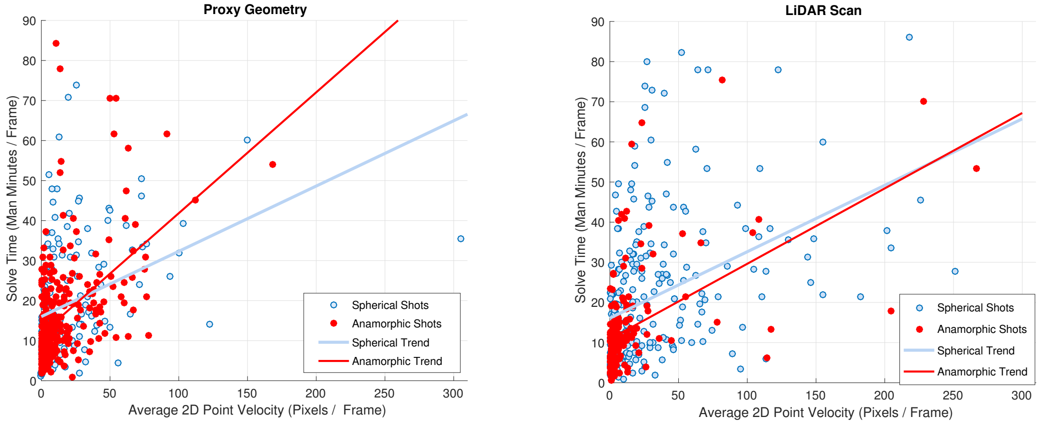


Figure 7: *Solve Time vs. Point Velocity for Different Types of Lenses at Different Levels of 3D Scene Data Availability. If anamorphic shots take longer to solve for than spherical lenses at the same velocity, the solid red line in each chart should have a steeper gradient than the blue line.*

4.4 Camera Constraints

In order to determine the constraints on the camera we analyse the absolute position curves for the solved 3D camera. In the case of a pure rotation around the camera's nodal point, we would expect the cameras translation to remain constant over time. For this work, shots will be characterised into the following 2 categories:

- *Pure Rotation* or a *'Nodal'* shot - a shot consisting solely of rotation around the camera's nodal point.
- *Free Move* A shot in which there are no constraints on the cameras movement

At present we have not developed a robust method of determining parallax in a scene using the data we have available to us in this dataset and we recognise that this will be a useful and important characteristic to analyse. Work into this area is ongoing and it is our intention to address this separately in a future work. This would be something to consider when interpreting the results of this test. Pure nodal shots will not exhibit parallax in a scene. However, freemove shots will have varying degrees of parallax present.

Due to the relatively rare occurrence of nodal shots in our dataset, 122 out of 939 shots, it was not possible to split the dataset into Nodal vs. Non Nodal for each combination of lenses and 3D survey data. Therefore, shown in Fig. 8 are the Solve Times vs. 2D Point Velocity for all shots grouped into Nodal vs. Non-Nodal motion. It can be seen from the gradient and the trend line that Nodal solves do tend to take a shorter amount of time to solve for, over all speeds of camera movement. Work to increase the reliability of this data, by taking into account the amount of parallax in the scene, is ongoing.

5 Discussion

In this paper, we have attempted to statistically analyse the process of matchmoving using information drawn from experiences of members of the matchmove department at Double Negative. Our quantitative data used was actual production data from recent Hollywood film projects. One of the major limitations of this approach has been that we were unable to control for individual factors in our analysis of the impact of various attributes on solve times. There are also likely to be biases in our dataset. Shots we attempt to identify as 'difficult' will likely have had steps taken to minimise their solve time prior to the solve being completed. Shots that are deemed to be complex by the show supervisor will most likely be allocated to a more experienced matchmove artists to solve, which would likely reduce the impact of difficulty on the solve time. If it is known ahead of time what the shot will consist of, it is also more likely

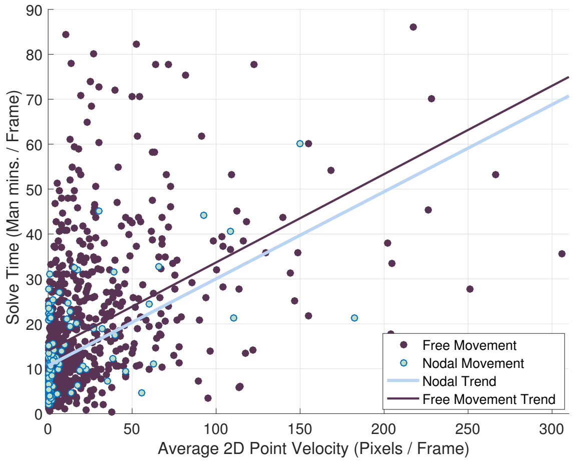


Figure 8: *Solve Time Vs. Point Velocity for Nodal and Free Move. Nodal Movement would be expected to take less time to solve for, and should lead to the blue line having a shallower gradient than that of the red line.*

to have had LiDAR or other on set data gathered. The generally higher point velocities found in shots where more 3D scene data is available, as shown in Fig.6, would suggest that this targeted data-collection is taking place.

Even after accounting for the impact of lens type, camera motion constraints and level of survey data available it can be seen that the variance of solve times for shots with similar velocity is large. For example, in figure 7 it can be seen that for a shot with a point velocity of approximately 50 pixels / frame, with full LiDAR scan available, and shot with a spherical lens, the solve time per frame varies from 10 minutes per frame to over 80 minutes per frame. This variance is however lower for a nodal shot of the same velocity, with solve times ranging from approximately 10 to 30 minutes per frame (shown in fig. 8). These values would therefore suggest that these factors alone do not give an accurate prediction for the time taken to solve a camera track. Furthermore, in general and across all of our dataset, the majority of shots analysed tend to cluster around the lower point velocities - suggesting slower camera movements are more common. This is to be expected, as only a few shots in most films will contain fast moving cameras (for example, during action sequences). This would also suggest that 2D point velocity alone is not an accurate method for predicting solve time.

This work does show that one of the most useful pieces of information available for speeding up the process of matchmoving is accurate 3D scene information to register 2D features to. Our quantitative results show this to be true over a large number of real shots, with a wide variety of camera speeds and using anamorphic and spherical lenses. Despite being a very active area of research, fully automated camera tracking solutions are not routinely used in the matchmove process. The results of our investigation would suggest that for shots with the same velocity of 2D image points (brought about by camera speed) having a LiDAR scan of a scene would allow for a solve to be completed approximately 10% faster than proxy geometry, and 20% faster than using 2D tracks alone. The dataset for shots with a LiDAR scan available has the largest occurrence of the highest levels of 2D point velocity (greater than 150 pixels / Frame), which would generally indicate a more difficult shot to solve. It can be concluded from this that LiDAR scans are currently regularly taken if it is known (from a script or VFX brief for example) that a shot might be challenging to solve. This would imply that this is considered a cost effective and accurate way of dealing with difficult shots, and more effective than the use of fully automated tracking methods.

This work has shown that the velocity of sparse 2D feature tracks do give some indication of the amount of time likely to be required to solve a shot, irrespective of the lens used, or camera motion constraint. We suggest that this could be a good use for automated 2D tracking methods, as an estimate for 2D point velocity does not require the camera to be scaled or lined-up to a 3D scene, which is a crucial part of the matchmove process.

At a more fundamental level, our experiences in producing this work have shown that there is value to the VFX industry in analysing production data to gain insights into processes in different stages of the pipeline. By performing relatively simple and computationally cheap analysis of solved shots - we have been able to determine the most common types of shot that we encounter (based on lens type, motion constraint and speed of camera movement), and the ways in which they can be solved quickly, along with causes for delays in the process.

Conclusions and Future work

We suggest the use of 2D point velocity obtained from sparse 2D feature tracks as a method for indicating the solve time of a shot. Our discussions with matchmove artists have highlighted the diversity of the work handled by matchmove departments in visual effects - and also the variety of methods used to solve for camera movement in these conditions. As it stands, the state of the art in computer vision methods for estimating camera movement fully automatically are not considered as reliable and efficient as manually solving a camera's motion using a combination of 2D tracking and additional scene information. Our quantitative results suggest that gathering 3D scene measurements is one of the most effective, and preferred, ways in reducing the amount of time spent on the camera tracking process in VFX work. In this paper, we have attempted to gather and analyse quantitative data from from real production footage. One of the biggest challenges with this approach is normalising and controlling for various independent factors. For example, we have been unable to control for levels of artist skill or experience for each shot analysed. We know that it is likely, in order to meet production schedules, that certain artists will be assigned shots that are more suited to their abilities and this will likely bias the results we obtain here. Commercial pressures mean that it would be infeasible to eliminate this bias from a live production for the purposes of statistical investigation. One potential method to determine the likely impact of this could be to create a synthetic sequence with absolute known camera movements, and examine the time taken to produce a solution over a group of artists with varying levels of experience. Although velocity appears to give an indication of shot solve time, the variance of shot times for 2D point velocities shown in Sec. 4 suggests that other factors might exist which might also be good indicators. Identifying these and the most reliable combinations of factors for shot time prediction is ongoing.

References

ARRI. Lens data system lds.

AUTODESK. Matchmover.

BAKER, S., AND MATTHEWS, I. 2004. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vision 56*, 3 (Feb.), 221–255.

BAY, H., ESS, A., TUYTELAARS, T., AND VAN GOOL, L. 2008. Speeded-up robust features (surf). *Comput. Vis. Image Underst. 110*, 3 (June), 346–359.

Bregler, C., Bhat, K., Saltzman, J., and Allen, B. 2009. Ilm's multitrack: A new visual tracking framework for high-end vfx production. In *SIGGRAPH 2009: Talks*, ACM, New York, NY, USA, SIGGRAPH '09, 29:1–29:1.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, IEEE Computer Society, Washington, DC, USA, CVPR '05, 886–893.

Dobbert, T. 2013. *Mathcmoving The Invisible Art of Camera Tracking*.

Fitzgibbon, A., and Zisserman, A. 1998. Automatic camera recovery for closed or open image sequences. In *Computer Vision ECCV'98*, H. Burkhardt and B. Neumann, Eds., vol. 1406 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 311–326.

Goulekas, K. 2010. *The VES Handbook of Visual Effects*. Elseveir.

Hartley, R. I., and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*, second ed. Cambridge University Press, ISBN: 0521540518.

Horn, B. K., and Schunck, B. G. 1981. Determining optical flow. *Artificial Intelligence 17*, 1, 185 – 203.

Hornung, E. 2010. *The Art and Technique of Matchmoving: Solutions for the VFX Artist*. Focal Press.

Huang, Y., and Essa, I. 2005. Tracking multiple objects through occlusions. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 1051–1058 vol. 2.

Jin, H., Favaro, P., and Cipolla, R. 2005. Visual tracking in the presence of motion blur. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 18–25 vol. 2.

Klein, G., and Drummond, T. 2004. Tightly integrated sensor fusion for robust visual tracking. *Image and Vision Computing 22*, 10 (September), 769–776.

Lowe, D. 1991. Fitting parameterized three-dimensional models to images. *IEEE Tra* (May).

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov.), 91–110.

Okatani, T., and Deguchi, K. 2002. Robust estimation of camera translation between two images using a camera with a 3d orientation sensor. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 1, 275–278 vol.1.

Rosten, E., Porter, R., and Drummond, T. 2010. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell. 32*, 1 (Jan.), 105–119.

Science D. Visions. 3dequalizer.

Shi, J., and Tomasi, C. 1994. Good features to track. 593–600.

Tola, E., Lepetit, V., and Fua, P. 2010. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 5 (May), 815–830.

Torr, P., Fitzgibbon, A., and Zisserman, A. 1998. Maintaining multiple motion model hypotheses over many views to recover matching and structure. In *SIXTH INTERNATIONAL CONFERENCE ON COMPUTER VISION*, IEEE Comp Soc, 485–491. 6th International Conference on Computer Vision, BOMBAY, INDIA, JAN 04-07, 1998.

Tu, Z., Poppe, R., and Veltkamp, R. 2015. Estimating accurate optical flow in the presence of motion blur. *Journal of Electronic Imaging 24*, 5, 053018.

Vicon. Boujou.

Xue, T., Rubinstein, M., Liu, C., and Freeman, W. T. 2015. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG) 34*, 4, 79.

Zhang, L., Portz, T., and Jiang, H. 2012. Optical flow in the presence of spatially-varying motion blur. *2013 IEEE Conference on Computer Vision and Pattern Recognition 0*, 1752–1759.

Zhang, Z. 2000. A flexible new technique for camera calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.