Focus Tracking for Cinematography

Aurel Wildfellner

Johannes Kepler University Linz

March 21, 2012

Contents

1 Introduction 1
  1.1 Motivation ................................................. 1
  1.2 Autofocus .................................................. 2
  1.3 Techniques in Cinematography ............................. 2
  1.4 Point Clouds ................................................ 4

2 Method 4
  2.1 Prototype Rig and Integration .............................. 4
  2.2 Viewfinder and Autofocus .................................. 6
  2.3 ICP Tracking and Reconstruction ......................... 6
  2.4 Strategies .................................................. 7
  2.5 Workflow .................................................... 8
  2.6 Hierarchical Sensor Fusion ................................. 8

3 Calibration 9
  3.1 Requirements and Challenges ............................. 9
  3.2 Intrinsic ...................................................... 9
  3.3 Extrinsic ...................................................... 10
    3.3.1 Extracting a Planar Marker ......................... 11
    3.3.2 Method for a DSLR Video Rig ....................... 12
  3.4 Focus Distance and Depth of Field ....................... 12
  3.5 Multiple Kinects .......................................... 14

4 Results and Evaluation 15
  4.1 Notes on General Equipment .............................. 15
  4.2 Panning Shot ............................................... 16
  4.3 Dolly Shot .................................................. 16
  4.4 Many Objects ............................................... 17
  4.5 Freehand Shot .............................................. 17
Abstract

Cinematographers primarily use manual control for focusing their cameras because existing autofocus techniques used in photography can’t be directly applied to video or motion-pictures and don’t provide sufficient artistic control. Adjusting focus therefore remains a key challenge, which limits the possibilities of executing certain shots. For instance, the amount of depth of field used in shots with a moving camera or subject is heavily influenced by how precise focus can be controlled. This work presents a simple method to overcome some of these challenges by tracking the focus with off the shelf sensory equipment and state of the art 3D point cloud processing techniques. The method integrates well with the current workflow of camera operators and their first assistants and even gives them more flexibility than a manually controlled follow focus. To evaluate the feasibility, a fully functional prototype was built and tested with professional camera operators.

1 Introduction

1.1 Motivation

Today, feasible and robust autofocus techniques for video cameras remain a challenge. Reasons are, that sensors can’t be integrated as easy as in still photography cameras like DSLRs and that algorithms and techniques used for continuous autofocus are not suitable for video applications. These and other reasons lead to the fact that professional camera operators, especially in cinematography, use manual control to focus a camera. Focus pulling, usually done by the first camera assistant or AC, requires a lot of skill and even if done highly professional, has it’s limitations. When planning a shot, one of the most important aspects is, if maintaining focus is possible at all and therefore often is a limiting factor.

With current technical advances these issues become more apparent. Cameras have become lighter, allowing the use of new rigs, which deliver more complex and faster movements. Consumer grade video cameras, which have a sensor of the size equivalent to Super 35mm film format or even larger than that with interchangeable lenses are available, which can deliver footage with extreme shallow depth of field. Therefore cinematographers gain more possibilities in executing extreme shots, but maintaining precise focus becomes difficult or even impossible. The goal of this project is to develop practical focus tracking and assist techniques, which enable cinematographers to execute shots otherwise being impossible. The key idea is, to use state of the art and emerging 3D vision and point cloud processing techniques to implement a practical and working system. With the release of the Microsoft Kinect this field of research has gained much attention lately and also existing techniques have become more practical to be used in engineering. With the newly developed Point Cloud Library (libpcl) [6] there is now a common framework for 3d vision and perception technology, which provides a good foundation for future research and development. Also faster GPGPUs with more memory give us the possibilities to use classical algorithms as RANSAC (random sample consensus) or ICP (iterative closest point) in real time applications. With all these technical advances it is now much more feasible to implement such a desired focusing system.
1.2 Autofocus

An autofocus uses a sensor, a motor and a control unit to focus a lens fully automatically or at a given metering point. An electronic rangefinder only measure the distance to an object without adjusting the lens automatically. There are two coarse categories, active and passive autofocus systems. Active means, that some sort of active sensing is done, by emitting sonar waves or light, while a passive autofocus only measures light emitted by objects in the scene.

Most comercial available cameras feature a passive autofocus. Modern DSLR cameras use a passive through the lens phase detection sensor, which is separate to the actual image sensor [20]. For this a beam splitter, integrated into the mirror, directs a portion of the light to the sensor and some to the optical viewfinder. When flipping the mirror the sensor is fully exposed and the sensor doesn’t receive any light. Therefore autofocus is not available while continuously capturing and exposing the image sensor in video mode. For video cameras it is also not desired and feasible to direct portions of the light to an separate sensor and not fully illuminate the image sensor.

Another passive autofocus technique is based on measuring the contrast of an image region. Most digital compact cameras use this, because no extra sensors are needed, since the picture from the main image sensor is analysed [10]. The focus is adjusted in a way, that contrast and therefore sharpness is optimized in some defined image region. At first, when a region gets out of focus it is not known in which direction and how far to adjust the focus. Therefore contrast based autofocus system feature complex search and adjustment strategies and algorithms. Some methods try to achieved high performance ad precision here, to make this technique also feasible for video cameras [11] [15]. However, searching for maximum contrast truly has it’s limitations in video applications and therefore isn’t much used in professional cameras.

There are also some systems which use a hybrid approach. One which is particularly interesting Fujifilm’s phase detection integrated into an ordinary image sensor by masking individual pixels [4]. In combination with contrast measurements, fast and precise focusing can be achieved without a separate phase detection sensor.

Continuous autofocus refers systems which continuously adjust focus to keep a moving object in focus. This is often achieved with different proprietary tracking algorithms and is featured by most commercially available cameras [7].

1.3 Techniques in Cinematography

Traditionally the first assistant cameraman, or focus puller, is controlling the focus of the camera, while the camera operator itself is responsible for framing and moving the camera. For this a so called follow focus device is used, as depicted in Figure 1. It gives better ergonomic control and also the ability to make marks for different positions for the focus ring. Transmission is done with a standardized pitched gear, which is usually present on cinematographic lenses. Often a motorized and remote controlled follow focus is used, so that the pulling focus doesn’t interfere with camera movement.

When planning a shot the focus puller has to be aware of the various ranges. For this, typically marks are used to guide the camera operator and actors to certain positions, for which the ranges were measured beforehand. Though this task seems rather uncomplicated, it takes a
There are some focus assist techniques, like external range finders, also called cine tapes, which can operate like a simple continuous autofocus [9]. For continuously adjusting focus at a fixed object there is also the possibility of encoding and measuring the position of a dolly and controlling focus with this measurement. But such electronic systems often take a lot time to be set up, so that manual control is mostly preferred.

Motion control refers to the process of driving movement of a camera electronically and motorized. Such systems can comprise computer controlled dollies, cranes or even robotic arms,
which give full six degrees of freedom. The movements of the camera are then set similar to keyframe animation and can be replayed automatically. By also controlling the lens, complex and fast focus pulling can be achieved. The downsides are, that motion control is rather complex and expensive, not suitable for many situations and also lacks flexibility when executing a shot.

A very popular rig, which allows a lot of freedom in moving the camera is the *steadicam* depicted in Figure 2. Pulling focus here is especially challenging and therefore shots are mostly limited to moderate deep focus [13].

Because of many technical advances in the last years, the work of a focus puller gets more challenging. The sensor size of video cameras constantly increase and with modern full-frame DSLR cameras, which can record video, even exceed classical 35mm film and motion picture cameras. This results in more shallow focus, which is also utilized by cinematographers. Another aspect is, that cinema-grade video cameras have become extremely light and small in the past years, which give filmmaker new capabilities in using and moving the camera in a more complex way. These developments make precise focusing even more difficult than it was with professional motion picture cameras a couple of years ago.

### 1.4 Point Clouds

Point clouds can be acquired from different sensors, for example:

- **Time of Flight Camera**: Resolve the distance based on the known speed of light.
- **Stereo Camera**: Can be used to compute a disparity map.
- **Kinect**: Is using an infrared laser projector with structured light and a camera.

The *Point Cloud Library* or libpcl [6] provides a framework to process such data. Many basic algorithms like for filtering and segmentation are implemented, but also higher level features as tracking and visualization. There are also many CUDA and real-time implementations of important basic algorithms as RANSAC (random sample consensus) or ICP (iterative closest point).

The current implementation of focus tracking relies heavily on libpcl.

### 2 Method

#### 2.1 Prototype Rig and Integration

To evaluate the developed techniques a fully functional prototype was built. For this, a Microsoft kinect is integrated into a DSLR video rig, depicted in Figure 3.

The rig is based on standard 15 mm camera support rods and clamps. The kinect is mounted rigidly to the rails and also removed from it’s plastic casing to reduce mechanical flex and physical size. The lens of the camera is placed very close to the kinect, the focal length is fixed and the entire camera is strapped onto an additional supporting bracket. The used camera is a *Canon EOS 600d* which can record video at full 1080p resolution and has a sensor format (APS-C) equivalent to *Super 35mm* film.
The used lens is a Canon EF-S 17-85mm f/4-5.6 tele. The focal length is set to 28mm which, considering the crop factor of the sensor, results in a field of view of a 42mm lens. Unfortunately the largest aperture at this focal length is f/4.2, which delivers a rather wide depth of field. Other lenses are still to be tried.

The focus is adjusted by an external RC digital servo motor (Futaba S9070SB). The internal motor of the lens wasn’t used, because there was no straight forward way to control it in real time and also because of lack in speed. Further external motors for lenses are very common and there are a lot of lenses worth using, which have no internal servo for focus adjustment. The servo position is set by a PWM signal generated by an Atmega microcontroller, which is put into the red plastic box located at the bottom rear of the rig. A gear is mounted to the servo, which transmits rotation to a geared belt strapped around the focus ring of the lens. This setup results in lots of backlash, but is still precise enough for a prove of concept.

The control box features a socket where a remote control can be attached. At the moment the camera cant be operated wireless, mainly because of the kinect requiring a full high speed USB2.0 connection. The rig, as is, weights about 3 kg, which could be easily decreased by using light rails and clamps. In practice no heavy equipment has to be used inherently.

In the front there is some sort of matte box, which serves as a combined lens hood for the DSLR and the kinect, as the lens hoods of the kinect were removed and the kinect makes no space to mount one to the large lens.
2.2 Viewfinder and Autofocus

The point cloud from the kinect is registered with the image of the DSLR resulting in a RGBZ image. Virtual meter points are defined in the image and can be selected by the operator. As with a classical autofocus system the camera can be pointed onto an object, a button is pressed, the distance is measured, the lens is adjusted accordingly and the object gets focused. The kinect serves as an external range finder, which is seamlessly integrated with the camera, requiring just one simple button as an interface. Canons default firmware doesn’t feature a good visual crosshair in it’s OSD, which could precise aiming. By using an alternative firmware [5] a custom crosshair can be displayed, which makes it easy to point the camera precisely on very small objects.

![Figure 4: Reconstructed point cloud of the scene. The white cloud represents the current view of the kinect and blue one cloud the range image registered with the DSLR camera. The lower coordinate frame at the left of the image marks the origin of the kinect and the upper one the DSLR. A ray is casted into the scene measuring a range of 1.76m (red sphere).](image)

The selected and to be focused 3d point is projected onto the viewing axis and the distance to the camera origin is evaluated. This assumes a planar focal plane, which is sufficient precise approximation for a typical lenses, which does not have an extreme wide angle.

The same integrated method which is used for focusing on a point in the scene is also used to select a static 3d point which is then tracked and kept in focus while moving the camera.

2.3 ICP Tracking and Reconstruction

With the current method static points in the scene can be kept in focus. Some strategies are implemented to select which of several tracked points is kept in focus at certain moment. This simple technique integrates well with the current workflow of a focus puller, which would usually just note relative distances between the camera and marked static points in the scene and then set the lens accordingly when executing the shot. By just tracking static points the AC can focus on specific points independently of the camera position and movement.

Rather than tracking single points or objects the entire 6DOF pose of the camera is estimated. For this a free implementation of KinectFusion [14] in libpcl [6] is used. KinectFusion is a real-
time camera tracking and reconstruction pipeline which is aimed for augmented reality applications. However, some of its properties and features make it especially interesting for the desired focus tracking application. First of all it provides extremely robust and practical SLAM (simultaneous localization and mapping), not present with many other existing techniques. In addition, the real-time surface reconstruction capabilities give additional possibilities. KinectFusion uses a real-time ICP implementation on a GPGPU to estimate the relative transformation between two subsequent frames, which is a pretty common algorithm for this task. Classical structure from motion or other SLAM systems rely on detecting sparse scene features for this. By operating and registering the entire dense 640x480 depth map from the kinect without any feature detection or downsampling a low of robustness can be gained. Each registered depth map is integrated into a volumetric representation of the surfaces in the scene called a TSDF volume. Rather than registering two subsequent raw depth maps or point clouds from the kinect, the current frame can be registered with the globally consistent and reconstructed surface model. This results in an even better and more robust tracking of the camera pose especially over a long period of time.

The TSDF volume can also be used for robust foreground/background separation, marking moving objects as outliers for the ICP algorithm. This way robust tracking is possible even with large moving objects in the scene, covering the bigger part in the depth map.

In future versions inertial sensing could be used as well. Because of scene reconstruction, foreground/background separation and robust outlier removal there is also the possibility to detect when the ICP algorithm converges badly and tracking fails. In such a case the system could fall back to inertial tracking using a gyroscope and accelerometer. Because of global reconstruction, the current point cloud from the kinect can be registered again with the scene after ICP tracking is working again. This way, short periods of inertial sensing don’t result in an accumulating error of the tracked 6DOF pose of the camera. This could be used for shots where for instance an actor walks by, covering the entire field of view, leaving no background for robust tracking.

2.4 Strategies

The operator can select multiple points when setting up a shot. The challenge now is how to select or use such points to determine the currently desired focus distance. Certainly this depends on the type of shot. As a first start only a few simple strategies were implemented. One always selects the closest point in respect to the middle point of the camera and directly sets focus without any delay resulting in a very fast focus pull.

Another strategy selects the two closest points in respect to the middle point in the video image and geometrically interpolates between both coordinates. This simple technique turned out to be useful in panning shots to perfectly synchronize a focus pull and a pan. It gives a very subtle effect to the focus pull, since it is sort of masked by the motion and motion blur, which makes the change in focus rather unobtrusive. A pan between a foreground and background object can be seen in Figure 5. The method also turned out to keep the focus while moving or panning the camera over long planes like walls of floors.

The current implementation turned out to be practically useful, but coming up with good strategies, algorithms and interaction techniques remains a problem to be solved in the future. It
Figure 5: The two red dots mark selected points. The turquoise lint is the viewing axis of the camera with the green dot marking the focal plane. The focus is interpolated between the two selected points while panning from one to the other.

is also interesting how to integrate more manual focus control and interaction into the system, like manually triggered rack focus between two tracked points.

2.5 Workflow

As previously stated the method integrates well in the existing Workflow of camera operators and their first assistants. A strategy is selected and several points are marked in the scene. It is also important to choose a proper initial point for the camera or register the reconstruction volume with the scene properly. At the moment the physical size of the volume is limited by a few meters (3*3*3) and therefore this needs some careful planning. The scene is partly reconstructed in the TSDF volume without any dynamic objects as actors and should comprise all static parts which will be present eventually.

In the actual recording phase the lens is then adjusted fully automatically, tho more manual interaction could be implemented at some point as well. These simple steps enable precise, fast and complex focusing scenarios.

2.6 Hierarchical Sensor Fusion

When selecting a 3d point in the scene some sort of depth map has to be evaluated at some pixel divining a virtual meter point. For this the raw point cloud from the kinect is registered with the camera image, or to be precise, a range image is splatted with the intrinsic and extrinsic parameters of the DSLR. Because of parallax errors, IR-light absorbing surfaces and too far ranges this depth map is not densely defined, leaving the possibility that the operator aims the
camera at a point where no range is defined. In such a case the system falls back and casts a ray into the TSDF volume, where chances are higher, that the point might be defined.

A reasonable addition to the system would be to also use a depth map acquired with a stereo camera. This could enable direct focusing on distant objects, which are out of the kinects operating range.

3 Calibration

3.1 Requirements and Challenges

The presented focus tracking method needs various calibration of cameras, sensors and mechanical parts. Calibration methods and tools used in classical computer vision scenarios usually face very different requirements and conditions and often deal with a much more static setup.

In a cinematographic scenario the used camera system is changed constantly, by exchanging lenses, using different camera support and rig setups and also by using entirely different cameras in the same production. A huge number of various lenses are used, which have different optical properties, than those favored for computer vision applications. These aspects lead to the fact that a time consuming and complex calibration procedure is not feasible. The challenge here is to use techniques, which enable a camera operator to calibrate the system in a couple of minutes and which trade precision for speed and ease of use.

3.2 Intrinsic

For focusing applications precise intrinsic calibration of the video camera is not necessary. When defining the principal point as a virtual metering point for measuring ranges and marking points in the scene, this is invariant to lens distortion and change in focal length. For the various tracking and focus pulling strategies only a rough registration of the point cloud with the video image is needed, which can be done by using the already known and approximated angle of view of a lens.

However, in general, for precise extrinsic calibration precise intrinsic calibration is necessary. Tho this is a time consuming process, it only has to be done once for each camera/sensor-format and lens pair. Even for a prime lens it is important to note, that the focal length will slightly vary when changing the focus distance. It is therefore necessary to calibrate intrinsic parameters for a known focus distance and then set the lens accordingly when doing extrinsic calibration. Usually someone prefers to calibrate intrinsic at a close distance, because handling markerboards is easier. Extrinsic calibration usually works with more distant markers. In practice it is rather easy to find a good balance for focus, by capturing images for intrinsic calibration at a moderate range with a small aperture.

All intrinsic calibrations were done with OpenCV and a chessboard marker. A usual number of 30-40 sample images were necessary for precise calibration. This was only done for a single focus distance. An extension for better intrinsic calibration when changing the focus distance would be, to do several intrinsic calibrations for an entire focus stack and interpolate values for the camera matrix and distortion coefficients. This would be necessary for very precise registration of the point cloud with the camera image, but would be laborious. Therefore it
is recommended to look into more efficient and automatable tools and techniques for intrinsic calibration before making such an attempt.

![Typical sample image for intrinsic calibration](image.png)

Figure 6: Typical sample image for intrinsic calibration. The chessboard corners have to be placed carefully at the borders of the image, to determine radial distortion correctly, which is a time-consuming process.

### 3.3 Extrinsic

To successfully register the point cloud of the kinect with the DSLR camera the relative transformation between both cameras has to calibrated (extrinsic). There exist lot of different approaches to this problem and some published methods seem to use rather unique methods [19]. The key issue is, that someone needs to correlate features in the point cloud with features in a plain RGB image. This can be done for e.g. by using a 2.5d pattern [16]. However, any method which relies on detecting small and discontinuous edges, like small holes in a surface, is not really suitable for the kinect itself.

Other methods use simple geometric objects, which can be easily detected in a point cloud and in 2d image. A sphere would be such a shape, the center can be easily found in a pointcloud with RANSAC and in a 2d image with a hough-transformation [12]. With this, pares of correlating 3d coordinates can be sampled and from this a transformation can be estimated. A slightly other approach is, instead of sampling pares of matching 3d coordinates, two trajectories of such a geometric shape are tracked, while it is moved around in front of both cameras. The two resulting point clouds can be aligned with ICP to compute the proper extrinsic transformation [18].

Another simple geometric shape which can be detected in a point cloud and in 2d is a plane, especially by using markers or chessboard patterns. Most of the so far mentioned techniques try to define point to point correlations. By detecting a plane this is possible too, for e.g. by extracting the center of a rectangle. However, a more accurate method is to just sample correlating plane rather than points and optimize a transformation, so that pares of planes align [8], [21]. The key challenge here is to formalize a good distance measurement between two planes, which is not as straight forward as the simple squared euclidean distance between two 3d points.

Currently the technique and matlab toolbox from [8] is used for extrinsic calibration.
3.3.1 Extracting a Planar Marker

Figure 7: Finding a planar marker in a point cloud. The turquoise sphere marks the estimated center of the rectangle.

For some extrinsic calibration techniques and in other situations it is necessary to extract a planar marker from a point cloud. The marker has a rectangular shape of a known size. It is mounted on a tripod in a way, that it the connection to the marker, doesn’t face the camera and is occluded by the marker itself. Therefore the markers respective points in the point cloud form an euclidean cluster and a marker can be extracted easily:

- Segment the point cloud into euclidean clusters. Reject clusters, which don’t fit a minimal number of points.
- For each cluster extract the largest plane with RANSAC.
- Project all the inliers of the fitted planar model onto the plane and check the dimensions. If the point cloud fits the known width and length the marker is successfully extracted.

To roughly check the width and length and to estimate the center of the marker, the diagonals are extracted:

- Select a random point.
- Find the point with the maximal euclidean distance. This converges to the point in one corner.
- For this corner point, again, find the point with the maximal euclidean distance. This converges to the point in the opposite corner. The diagonal is given.
- The other diagonal is found the same way by choosing a proper initial point based on the first diagonal.
- To estimate the center of the marker, the center between the center points of both diagonals is computed.
3.3.2 Method for a DSLR Video Rig

As previously mentioned, a video rig has to be constantly changed and modified, including remounting of the kinect, which changes its extrinsic in respect to the video camera. Registering a point cloud with a camera image is a time-consuming process and despite of many possible optimizations it requires many sample images from different poses.

However, the extrinsic between two ordinary camera images can be estimated easily, even by a single pair of images with one marker. The accuracy can be improved by sampling more images, but 4-5 seemed enough for its application. The kinect also features a RGB-webcam with a resolution of 1280x1024 pixels, which is rigidly mounted in respect to the point cloud sensor. The extrinsic between this camera and the point cloud origin is calibrated only once for each unit by detecting a planar marker in 3d and the corresponding chessboard marker in 2d. With the corresponding planes or center points the 6DOF transformation of the extrinsic is estimated.

When mounting the kinect to a rig, it is now sufficient to calibrate the extrinsic between the video camera and the internal webcam of the kinect. Multiplying the previously calibrated extrinsic, internal to the kinect, with this new extrinsic gives the correct transformation matrix to register the point cloud with the image of the video camera. As mentioned, taking four to five images of a chessboard marker is enough for this, which can be done in less than one minute. With these simple steps, fast and feasible extrinsic calibration for a film set scenario is possible.

3.4 Focus Distance and Depth of Field

To be able to focus on an object at a given distance, the lens focus ring position and the driving motor have to be calibrated. The result is a transfer function which maps a discrete motor or servo position to a focus distance.

Some high grade cinematographic lenses, but also some Canon EF lenses, feature encoders for their lens settings, so that the position of the focus and focal length ring can be recorded. In photography this is used to roughly determine the focal length for an EXIF tag of an image and measured focus distance is used for flash exposure compensation. Therefore the precision of those measurements is limited and varies a lot with different lenses [2].

In commercial available lens control systems, as ARRIs WRS [1], lens data can be recorded and archived with time-code data. A servo motor, which drives the lens, is then turned step by step and the according focus distance is recorded, thus calibrating the transfer function. The AC can then set the focus distance directly by a value in meters.

To calibrate the focus distance of the introduced prototype, such a simple method was also implemented. The used lens is a Canon EF-S 17-85mm f/4-5.6, which delivers relative precise readings of the focus setting. By installing a third party firmware [5] on the EOS 600d, it is possible to get access to this value (Figure 8). The servo is then stepped through all possible positions and the according focus distance is recorded, thus calibrating the transfer function. However many lenses don’t provide accurate or even no electronic reading of the focus distance. For its application a very fast (f1.2 - f1.4) and compact lens is desired, which further limits the number of possible choices. Therefore another way to measure the focus distance of those lenses is required. This was done with a contrast-based autofocus and by measuring the
distance towards a testpattern with the calibrated kinect sensor.

Figure 9 depicts a possible setup for calibrating the focus distance for a lens drive motor. A testpattern is placed on a linear slide. The idea here is to align the axis of this slide with the optical axis of the camera. For this the carriage is first placed at the front of the slide and a aperture, a black film with a tiny hole, is aligned with the center of the testpattern. The carriage is then moved to the back of the slide and the camera is focused onto the testpattern. The camera is now carefully placed, such that the center of the testpattern is visible through the tiny hole in the aperture, which results in an alignment of the optical axis with the axis of the linear slide.

The testpattern is then placed at some distance, the camera is focused, the distance towards the pattern is measured with the kinect and the value of servo-position is recorded. Several such samples are taken. The kinect is limited to measure ranges exceeding 0.5 meters. When the testpattern is moved closer to the lens, the distance is measured relative to a previous sample with a measuring tape stucked to the floor (this is also the reason why exact alignment was done). A transfer function can be estimated by fitting for e.g. a polynomial function to the recorded samples, depicted in Figure 10.

To automate the focusing of the camera in this calibration process a contrast-based autofocus tool was implemented. Images are is captured from the 600d from it’s analog composite video port in PAL resolution, which delivers sufficient quality when using additional digital zoom. The tool *contrastaf* analysis these images for contrast and finds the optimal focus setting.

```
# $ ./contrastaf -w 0.4 -d /dev/video0 -m /dev/ttyUSB0 -v
```

Focusing with servo position: 118

A possible extension for the future would be to consider backlash for different directions resulting in two different transfer functions for clockwise and counter clockwise adjustments of the focus mechanics.
3.5 Multiple Kinects

In the prototype rig the kinect and the video camera face the same direction and the entire field of view of the video image is covered by the depth image. This is partly necessary, so that the kinect can be used as an external range finder. However, after some first practical tests in a studio environment an issue became evidently very quickly. In some shots the framing of the camera
and therefore also the captured point cloud is highly suboptimal for tracking the pose and failed quite often. For instance, if only an actor in front of greenscreen is visible, two subsequently captured point clouds can’t be robustly registered, because a moving actor would only result in outliers and the background, a simple plane has no features for robust alignment.

A practical and straightforward solution to this would be, to use two separate kinect sensors, one as an external range finder, where the point cloud is registered onto the video image and one for the actual 6DOF tracking. The one responsible for tracking could face a different direction, capturing structures, which enable more robust point cloud registration. It could face the ceiling (in studios usually with lots of lighting equipment), the ground (dolly tracks) or the side, always being rigidly mounted in respect to the video camera and the other kinect.

However calibrating the extrinsic for a kinect mounted in such a way and the video camera is not as straightforward as in the usual setup, because their field of views don’t overlap at all. Therefore extrinsic can’t be directly estimated by finding corresponding features in the point cloud and the video image.

A more practical approach is, to calibrate the extrinsic between the two kinects and the extrinsic between the kinect and aligned video camera and compose both extrinsic transformations. Some analysis of fusing and calibrating multiple depth images is discussed in [17] and [12], but only considering the case of overlapping field of views.

A possible solution to this problem is to use a combination of scene reconstruction and point cloud registration. An important aspect, as previously mentioned, is to be able to to calibrate the extrinsic within a few minutes. Some first tests were done by using the kinect fusion implementation [14] of libpcl. The camera is rolled and tilted in a way, that both depth cameras are leveled. The entire rig is then panned 360 degrees to capture a panorama and reconstructing a point cloud for each kinect. These two point clouds are then registered with ICP (iterative closest point) to get the transformation between the two kinects. How such a registration technique can be used for extrinsic calibration is discussed in by Makadia et al. [18]. A known problem for ICP registration is, that the point clouds need to have some initial alignment. This can be done for e.g. by roughly pointing the rig at a marker with one kinect and then pan/tilt/roll so the other kinect is pointing at the marker and record this transformation to get some rough initial alignment pose.

4 Results and Evaluation

The prototype rig was tested and evaluated for several standard setups and shots in a studio environment with film and media students. A short summarization can be seen in this video [3]. The results were mostly as priorly expected and showed the systems potential but also the current limitations. All of the following discussed shots can be seen in the video.

4.1 Notes on General Equipment

It may seem insignificant for a prototype in this stage, but the required cabling posed a severe problem and therefore limits proper evaluation. Currently the rig needs a lot of external cabling, which could be partly replaced by batteries and radio control. However, the large bandwidth
required by the kinect, which needs about full USB 2.0 speed (480 MBit/s), can not be easily transmitted wirelessly. Some wireless usb (WUSB) repeater could solve this issue here, but may lack linux support etc. (to be investigated). At the moment this also makes the use of a steadicam impossible, where no external cabling can be used at all.

Size and weight seem to be fine, tho could be improved a lot. But camera operators seem to have no problem with a 3 kg rig anyway.

4.2 Panning Shot

A panning shot, where the camera remains on a static tripod is pretty straight forward. We had nearly no problems with the prototype here and it already worked very smoothly. The operator picked two points and paned between them. It is often desired to hide a focus pull in a pan, so it is hardly recognized and has a subtle effect. The strategies for how focus gets adjusted must be improved here, but by simply snapping the focus to the correct distance when the objects enters the field of view already results in a precise and inconspicuous focus pull.

4.3 Dolly Shot

We carried out several dolly shots where we move the camera towards an object. The usual workflow would be to make several marks on the dolly track as a guidance. Setting this up for an AC takes some time and also using an electronic range finder is cumbersome in such a situation. While there are many other ways to control focus precisely in such a shot, they are challenging, but feasible. However, focus tracking needs hardly any time to been setup here and provides nearly perfect focus control.

Figure 11: Continuous focus on a small object in a dolly shot. Working title: nutshot.

In one shot we focused on a hanging walnut, a rather small object, and moved the camera towards depicted in Figure 4.3. With a classical external rangefinder it would not be possible to track such a small object, so another larger object has to be tracked for reference, which complicates things further and takes more time for setup.

Our findings were, that especially the saving of time for such kind of shots are a big advantage.
4.4 Many Objects

Also a mixed dolly/panning shot with the camera resiging on a small crane mounted on a dolly was performed. Many alternating foreground/background objects were brought in focus. Doing this precisely would be also a challenge in practice, because the grip moving the dolly and the operator panning the camera have to be perfectly sync and an AC would need measure the according ranges and set them at the right moment. In practice there is a limit in how precise these three people can coordinate and execute a shot in sync. With focus tracking executing such a shot was pretty easy and straightforward.

4.5 Freehand Shot

The prototype was placed on a shoulder rig (Figure 12 to be freely carried around. Here the limitations of tracking and the small reconstruction area got apparent. The scene was relatively wide and open, leaving a few references for stable tracking. Often the background was too far away, to fit in the reconstruction area of the voxel grid. After many failures we managed to get some good results of precise and complex focus control. The operator was moving forward towards an object keeping it in focus. Then he moved back and panned over to a standing actor. Such a situation would have been extremely difficult for a focus puller, especially because the operator was freely moving without any fixed marks or references.

Figure 12: Prototype on a shoulder rig.

If the tracking is further improved as partly mentioned and can also cover a physical larger scene, the system holds a lot of potential for such kind of shots. The next logical step here is, to do this on a steadicam.
References


[12] Li Guan and Marc Pollefeys. A unified approach to calibrate a network of camcorders and tof cameras.


