BIRS Workshop on
Mathematical Methods in Computer Vision

30 Sept – 5 Oct 2006
Banff International Research Station
Alberta, Canada

Organizers

Martin Jagersand (U. Alberta)
Anders Heyden (U. Malmö)
Peter Sturm (INRIA)
Bill Triggs (CNRS)
Dana Cobzas (U. Alberta)
Jim Little (U. British Columbia)
Steve Zucker (Yale)

`http://www.cs.ualberta.ca/~vis/vision06`

1 Introduction

There is a long and fruitful relationship between mathematics and image understanding starting with the desire to understand perspective in renaissance painting and art. This trend continued with the development of non-Euclidean geometries over the past two centuries, and it has expanded significantly over the past five decades with the introduction of new 3D digital imaging systems in many scientific fields and the desire to employ machine vision in applications ranging from robotics to medical imaging. Computer Vision started as a subfield of AI, but it has expanded into a broad field using a wide range of methods from mathematics and statistics. The formal study of image formation in the framework of projective geometry has produced many new methods for uncalibrated cameras, visual correspondence and visual invariants over the past twenty years, more recently expanding to cover very general camera models and algebraic geometry based methods for studying critical configurations. Another very active topic is the use of variational methods and PDE's to compute image segmentations and 3D surface models. The BIRS *Mathematical Methods in Computer Vision* workshop sought to span a broad range of the applications of mathematics to computer vision, from newly emerging research directions to mathematically-motivated algorithms that are have become practical enough to find their way into applications. Finally, a lab session organized jointly with the Banff New Media Institute went full circle to explore the use of modern computer vision methods in the creative process for example for the creation of models and animations.

Loosely speaking, computer vision is the study of how to compute properties of the 3D world from 2D images. One major line of study takes this rather literally, focusing on computing a representation of the 3D geometry and appearance that is as complete as possible given the 2D

images. Three areas related to this approach were covered in the workshop: geometric vision; variational methods for estimating surfaces and appearances; and applications of scene reconstruction. While a complete representation is useful for many engineering and entertainment applications, notably ones involving graphical rendering, there is increasing evidence that it is not necessary for many biologically relevant vision tasks. Much useful information can be derived from the 2D visual signal without the recovery of explicit 3D information. Two areas where the workshop represented this line of research were learning for visual recognition, which is often performed using only 2D image signatures, and human motion tracking, which can take place in 2D or 3D but which often involves 2D models.

In addition to technical presentations on the above topics, we singled out several areas that we felt were of interest to a broader audience for special interaction sessions. These included both generally accessible talks and hands-on interaction through demos and posters. One set of sessions was aimed at the general public, including talks and demos of photo album organization in 3D "photo-tourist" and scanning and digital display of virtual heritage. For details see the summaries below. A second set of interaction sessions was designed to encourage collaboration with the Banff Arts centre. The session was attended mainly by Banff New Media Institute modelers and animators. It contained talks and hands-on demos on 3D modeling from 2D images of scenes and objects. The participants started with real physical objects or with their own images and by the end of the interaction they had computed the corresponding 3D digital models. Another interaction topic was projector guided painting, which showed how computer vision can analyze and aid novice painters. Again, more detailed summaries are given below. To complement this, the Banff New Media Institute offered a visit to its lab that included a tour of the facility and an in-depth demo of its 3D visualization "cave" and its modeling and animation studios.

2 Presentation Highlights

2.1 Global Optimization and Large-Scale Problems in Geometric Computer Vision

Fredrik Kahl of Lund University (Sweden) presented a framework for approaching globally optimal solutions of geometric computer vision problems (slides: PDF, PPT). The overall goal is to find the best model that is consistent with the observations. In the context of geometric computer vision, this means that the differences between the reprojected 3D scene and camera geometry and the image measurements should be minimized. Traditional reconstruction algorithms use local descent and often fail owing to local minima.

Fredrik Kahl presented approaches that guarantee to find a solution with a residual error within any desired tolerance (if such a reconstruction exists). His first approach is in the Branch & Bound class and uses convex envelopes whereas his second uses lower-bounding relaxations based on Linear Matrix Inequalities (LMI). The approaches are applicable for squared or absolute residual differences (the latter being a more robust norm). He showed how to apply them to several typical geometric computer vision problems including n-view triangulation and space resection (estimation of intrinsic and extrinsic camera parameters).

Richard Hartley of the Australian National University described properties of quasi-convex optimization problems and how to use them in computer vision problems. When using the L_∞ norm to account for differences in image measurements and reprojected 3D scene and camera geometry, many common geometric computer vision problems turn out to be quasi-convex. Various properties of quasi-convex functions that make their global optimization feasible were described.

In practice, one is usually confronted with a certain percentage of outliers in the image measurements. These are typically dealt with using "RANSAC" type approaches in which initial

solutions are computed from random *minimal* samples of measurements and checked for consensus with the remaining measurements, until a satisfying solution is found. Richard Hartley showed that for quasi-convex problems, the global solution with respect to *all* measurements has the property that, if the measurements contain outliers, there is at least one among the measurements with largest residual. This could be the basis for deterministic estimation procedures that guarantee to find all of the outliers in the measurements and thus the globally optimal solution of the considered estimation problem.

Ananth Ranganathan of Georgia Tech presented solutions for performing inference on large-scale graphical models and applied this to matching problems in geometric computer vision. (Slides: PDF, PPT). The main motivation is to be able to handle large-scale reconstruction problems where appearance information is unreliable or unavailable, thus posing challenges for image matching. This is the case in the 4D-Cities project at Georgia Tech (modeling of entire cities in 3D and over time using photographs or videos) as well as in laser-based Simultaneous Localization and Mapping (SLAM) in robotics. In such cases, reliable matching can be approached by computing marginal covariances for the structure and motion variables and using these to perform maximum likelihood correspondences.

It is well-known that such marginal covariances can be expensive to obtain. Ananth Ranganathan explained how to formulate geometric computer vision and SLAM problems using graphical models so that inference on them can be explained in a purely graphical manner via the concept of variable elimination. This leads to a new way of looking at inference that is equivalent to the junction tree algorithm. When applied to linear(ized) Gaussian problems, the algorithm yields the familiar QR and Cholesky factorization algorithms. This connection with linear algebra in turn leads to strategies for very fast inference in arbitrary graphs, such as those encountered in the above mentioned computer vision and robotics problems.

2.2 Multi-view Geometry

David Nistér of the University of Kentucky discussed the problem of determining the relative locations of a set of microphones, simply by recording unknown sound sources. More precisely, he assumed that "correspondence" is solved, so that time-difference-of-arrival (TDOA) measurements are available for each pair of microphones and that sound sources are distant (they are modeled as being located infinitely far away). David Nistér showed that under these assumptions, the location of a set of microphones and sound sources can be computed via an elegant matrix factorization. He also discussed degenerate and minimal cases. For example, to locate four microphones, at least six sound sources are required. The proposed formulation offers many similarities to concepts used in multi-*view* geometry, such as the absolute conic and multi-linear matching constraints.

Anders Heyden of Malmö University presented a framework for unifying discrete and continuous approaches for camera motion estimation (slides: PDF, PPT). Traditionally, motion estimation has been approached either from a pure discrete point of view, using multi-view tensors, or from a pure continuous point of view, using optical flow. Anders Heyden showed how to unify the two and derive hybrid methods combining the best part of each of them. This is embodied by differential-algebraic matching constraints that can be used for handling motion estimation in mixed scenarios containing both widely separated and closely spaced cameras. He also showed how to update the motion parameters from image correspondences, requiring fewer points than the traditional methods and also avoiding the non-linear constraints that usually appear in the calibrated case. Finally, he presented extensions to trifocal tensor[1] based motion tracking of a rigid stereo-head.

[1]The trifocal tensor is an algebraic constraint linking corresponding points in three views.

2.3 Segmentation of Dynamic Scenes

Marc Pollefeys of the University of North Carolina presented two families of approaches to obtaining 3D reconstructions of dynamic scenes, using one or more cameras. (Slides: PDF, PPT). The first set of approaches concerns the analysis and recovery of articulated motion with non-rigid parts, e.g. human body motion with non-rigid facial motion. The motion of points on the observed surface is modeled using a set of intersecting subspaces. By adopting an affine projection model for the camera, the observed motion can be analyzed and recovered using subspace methods. Overall, the approach allows motion segmentation to be performed to recover the underlying kinematic chains and object shape. An example is shown in figure 1.

The second focus of Marc Pollefeys' talk was on recovering dynamic shapes from silhouettes extracted in image sequences acquired with multiple cameras. This is a standard approach in computer vision, but it is usually assumed that throughout the image sequences, the objects of interest are entirely in front of the background, which is static. Marc Pollefeys addressed the case where the objects may be temporarily occluded by static objects during their displacements. He developed a probabilistic formulation for recovering the occlusion patterns and the dynamic object shape, as well as the shape of the occluding objects.

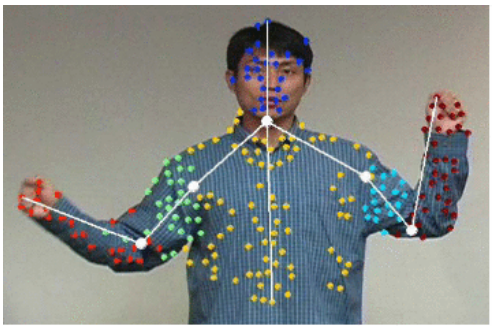

Figure 1: An example of motion segmentation and kinematic chain construction obtained with the approach presented by Marc Pollefeys.

Rene Vidal of Johns Hopkins University presented a framework for segmenting scenes containing independently moving objects and/or dynamic textures. (Slides: PDF). An example of such a scene is a bird floating on water: the bird moves independently from the rest of the scene and the appearance of the water is continuously changing. One can model such scenes as the output of a collection of dynamical models exhibiting discontinuous behavior both in space, due to the presence of multiple moving objects, and in time, due to the appearance and disappearance of objects. Segmentation of dynamic scenes is then equivalent to the identification of this mixture of dynamical models from the image data. The difficulty is that although the identification of a single dynamical model is a well understood problem, the identification of multiple hybrid dynamical models is not: in order to estimate a mixture of models one needs to first segment the data and in order to segment the data one needs to know the model parameters. Vidal proposed to approach this "chicken-and-egg" problem using a technique called Generalized Principal Component Analysis (GPCA). In the case of data living in a collection of subspaces, this proceeds in two steps. First, a general vector of polynomials is fitted to all of the data points (without segmentation); the polynomials represent a meta-model whose coefficients are the tensor products of the coefficients of the lower-degree polynomials representing the individual component models. The parameters of individual models can be found by differentiating

the fitted polynomials. Rene Vidal showed how to apply this framework to diverse problems in computer vision, such as image/video segmentation, 3-D motion segmentation, dynamic texture segmentation, and heart motion analysis.

Raghav Subbarao of Rutgers University explained how to generalize non-parametric estimation to data belonging to differential manifolds rather than Euclidean spaces, with applications in computer vision. (Slides: PDF, PPT). Many computer vision tasks involve the parameter estimation in the presence of noise and outliers. An alternative to parametric model fitting is the use of non-parametric techniques such as the popular mean shift mode discovery algorithm. For example mean shift can be used for clustering data points from some feature space representing visual cues, and also for image segmentation, object tracking, image smoothing, etc. Previous computer vision applications of mean shift have assumed that the data points belong to vector spaces but in reality the geometric constraints involved and the nature of the imaging device often lead to non-vectorial feature spaces. This is the case for example when the goal is to estimate one or several rigid motions between pairs of images or point sets. In such cases the feature space often still exhibits the regular geometry of an analytical manifold. Raghav Subbarao developed a nonlinear mean shift algorithm that generalizes Euclidean mean shift to analytic manifolds. As examples he considered two frequently occurring classes of parameter spaces, Grassmann manifolds and matrix Lie groups, using the algorithm for motion segmentation, model-based optical flow field segmentation and diffusion tensor based image smoothing.

2.4 Scene Reconstruction I

Yasutaka Furukawa of the University of Illinois presented a multi-view stereo algorithm that reconstructs a scene as a dense set of patches, i.e. points with associated surface normals. (Slides: PDF, PPT). The algorithm has two main steps. First, sets of feature points are detected in each image and matched across multiple images. Each match yields a single patch; the generated patches are sparse and only correspond to regions with salient image features. To densify the coverage, the second pass of the algorithm expands the set of matches, recursively adding new patches in the vicinity of existing ones. These are initialized based on the location and orientation of an existing patch then refined by optimizing a photo-consistency measure over patch orientation and depth relative to a reference image. The proposed method is different from existing multi-view stereo algorithms in various ways. First, it does not require any initialization, such as the visual hulls or bounding boxes that are often used to start the iterative deformation of surface based models. Secondly, it includes only a very small amount of regularization; this may be a drawback in some cases but it also allows the method to handle objects with complicated topologies. Thirdly, the method is memory efficient – the memory usage is proportional to the size of the inputs and outputs. It is also robust to outliers in input images such as moving pedestrians in an outdoor scene being reconstructed. Yasutaka Furukawa showed impressive experimental results on various data sets – for an example see figure 2 – along with a qualitative and quantitative comparison with state-of-the-art image-based modeling algorithms and laser range scanners.

Kyros Kutulakos of the University of Toronto presented Confocal Stereo, a novel approach for high-resolution 3D photography using a single camera. In contrast to other image-based approaches, the method allows depth maps to be computed using pixel-by-pixel processing, thus allowing the reconstruction of very fine surface details and scenes with multiple depth discontinuities occurring within several pixels (e.g. high-resolution images of hair). Confocal stereo works by taking a number of images of the same scene with different focus and lens aperture. At its heart is the confocal constancy property: as the aperture varies, the pixel intensity of a visible in-focus scene point varies in a predictable way that does not depend on the scene. To exploit this, Kutulakos developed a detailed lens model that factors out

Figure 2: One input image and the corresponding view of a 3D model reconstructed with the approach of Yasutaka Furukawa.

the geometric and radiometric distortions in high resolution SLR cameras with wide-aperture lenses. He showed how to recover the model from images and how to use it to reconstruct detailed 3D shape for a variety of complex scenes.

Gabriel Taubin of Brown University presented a 3D reconstruction method that exploits images taken with multiple flash exposures. (Slides: PDF). It extends the range of so-called shape-from-silhouette algorithms, which recover 3D shape based on object silhouettes extracted in images. Silhouettes correspond to the outer contours of objects and do not for example provide information of 3D shape within concavities of the object surface. The proposed approach uses sets of images. At each viewpoint several images are acquired under flashes in different positions. Depth discontinuities are detected by extracting the shadows they cause under the different flash-based illuminations. Geometric reasoning then allows the positions and orientations of points to be reconstructed, even when they are located inside concavities. However, points that do not produce observable depth discontinuities can not be recovered and so the method only produces a sparse and unevenly sampled representation of object shape. To handle this Gabriel Taubin described a method for fitting an implicit surface to the oriented point cloud, which is then used to generate additional oriented points on the surface of the object in regions of low sampling density. He then presented some 3D reconstructions of objects with or without texture and with rather fine surface details. He also also showed how to enrich the resulting geometric models with appearance information by fitting a Phong reflectance model to the observed image data.

2.5 Mathematics and Vision meets Arts: Banff Centre Interaction

Gabriel Taubin of Brown University presented a retrospective of pioneering work joint with Holly Rushmeier and Fausto Bernardini on the digital capture of Michelangelo's Pieta sculpture, covering technical, organizational and heritage research issues. At the time when this work was done real world capture projects outside the laboratory were a relatively new endeavor. For the project Gabriel's team developed hardware to capture the 3D geometry and an operational procedure to scan pieces of the object, merge the resulting collection of geometric meshes, and register texture images taken with a digital camera. Figure 3 illustrates the work flow. An important facet of the work is its handling of the numerous practicalities associated with real capture work in museums – an aspect rarely discussed in research papers but important in

bringing vision into real-world use. Close collaboration and a good understanding of museum rules and operation was required to make this collaboration between scientists and custodians of historical heritage a success. Gabriel's talk gave insights into many of these issues from the early stages of planning and securing access, to the practicalities of capturing a large object while working in a cramped space not designed for capture purposes. Finally, novel uses for the digital model were presented, such as being able to view parts of the statue from viewpoints otherwise occluded and the ability to speculate about alterations and modifications by trying them out on the digital model.

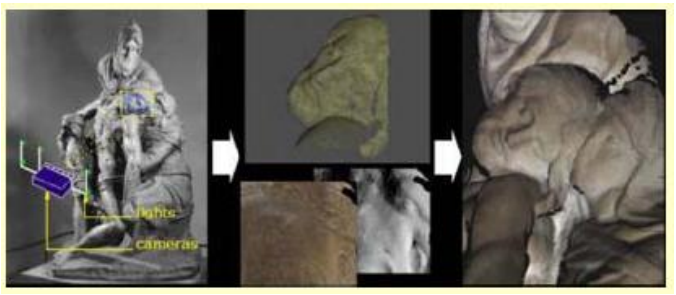

Figure 3: From real object to digital model through piecewise acquisition of geometry and registration of texture images.

Noah Snavely from University of Washington and **Richard Szeliski** from Microsoft presented a new way to represent and browse large collections of snapshots of scenes such as popular tourist sites. (Slides: PDF). The approach seeks a balance between the complexity and difficulty of acquiring 3D information and the richness of the representation displayed to the user. Much computer vision research seeks to create a complete texture mapped model of the scene, but to date the corresponding algorithms are not universally robust and it is usually impossible to guarantee the complete coverage of a complex scene, especially when details may be viewed at a wide range of scales. As an accessible middle ground, the authors propose to use a sparse geometric model as a means of registering the photos and camera positions in 3D and navigating the scene and the collection, while still basing visual browsing on the original photos. An example of the representation is shown in Fig. 4. Enhancements such as smooth transitions between views and the display of a sparse stylized geometry provide a satisfying 3D browsing experience. Several models based on unorganized photo collections from Internet photo sites were shown. The results are currently being developed commercially by Microsoft Live Labs.

Neil Birkbeck and **Adam Rachmielowski** from the University of Alberta talked about a three-level representation for capturing photo-realistic 3D models from 2D images. (Slides: PDF). On the macro scale, a conventional triangulated geometric model is captured using established shape-from-silhouette and structure-from-motion methods. At the level of texture capture, most previous approaches have attempted to work directly with the resulting triangulated models, but in practice they are often too coarse or inaccurate for good results. To improve on this, Neil presented a variational method that simultaneously refines the geometry and estimates a reflectance model. A novel aspect is that the representation in this step uses a depth map w.r.t. the model facets, thus avoiding the generation of a complex model with many small polygons while generalizing the image-plane disparity models typically used in conventional stereo. This representation also naturally supports recent displacement mapping methods thus providing more efficient rendering. Finally, on a micro level, instead of a conventional texture based representation of surface color, a linear basis much like those used in movie compression

Figure 4: On the left is an iconic representation of the Prague city square derived from photos that one of the presenters took during a conference trip. The estimated camera locations are shown as small rectangles. On the right is the view of one photo in the photo-tourist navigation interface, and surrounding the image a selection of other nearby (in a 3D sense) photos. Several navigation modes are available.

is used to encode the residual between the estimated model and the image inputs. This basis can be modulated and correctly interpolated by indexing it on viewing angle, thus allowing the rendering of arbitrary views not seen in the input images. Experiments illustrating the capture of both objects and people were shown – see Fig. 5 – and the models were integrated into a virtual city model inserted in the Edmonton landscape.

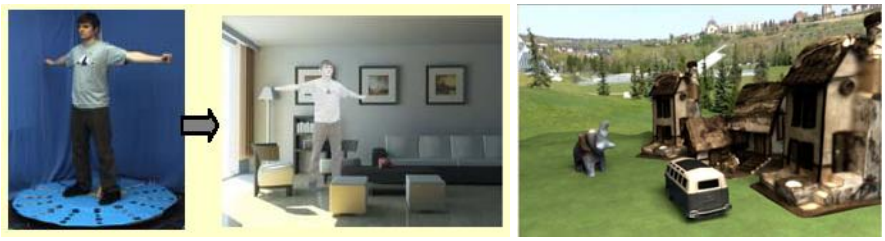

Figure 5: Object capture and insertion into a virtualized world.

Jim Rehg of Georgia Tech presented a system that trains novice painters using computer vision to control overlay projectors. (Slides: PDF). The system is designed to support the traditional painting experience and to intrude as little as possible on it. Rather than having to use digital input methods (electronic pens and tablets, digitizing devices etc), the budding artist uses conventional brushes, colors and canvasses. A computer vision system monitors the progress of the painting and provides appropriate augmentations via two computer projectors. A variety of painting aids and hints have been implemented and tested, from high level guidance to detail brush work. In teaching layering of paintings, the system uses the projectors to enhance or mask various areas, guiding the novice through the process of building the painting sequentially from the base layers up – see Fig. 6. There is also a visual aid for mixing colors, and a tone blending assistant to help with shading. To situate the current state within the time line from bare canvas to final painting, the system can alter the appearance of the canvas to show goals in a preview mode. It can also be completely turned off whenever the painter desires.

Figure 6: A projector based system for training novice painters. Left: the setup with two projectors and a video camera. Right: a projector-based augmentation showing paint layering.

The system has been tested by a number of painters, receiving positive feedback. Different users use the system in very different ways but one indication of overall usefulness is timing measurements that show that most participants choose to use the augmentation most of the time rather than painting unsupported with the system off. The talk generated a considerable amount of interest from the artists participating in the session. Both questions and criticisms were aired.

Adrian Broadhurst of Vicon talked about using commercial motion capture for creative applications and showed examples and clips of both the intermediate stages and the final results in movies. The Vicon motion capture system is based on retro-reflective markers and infrared lighting, see Fig. 7. Marker positions are detected by custom-made cameras with on-board processing, with the thresholded binary images being sent to the PC for further processing and 3D structure computations. Special care has to be taken to get the correspondences between different images of each marker correct. This is performed by using kinematic rigidity and temporal track consistency constraints. An XML representation is used to describe the character topology and joint types. From this a more detailed calibration is obtained by having the actors perform a range of test motions. Finally, tracks of the desired movements are recorded. A recent development in commercial motion capture is that multi-camera systems now can record the motions of several interacting actors at once, instead of having to record each actor separately. Adrian finished by showing clips from feature films where Vicon's systems have been used for special effects.

Figure 7: The Vicon motion capture system. Retro-reflective markers are fitted to the subject and tracked as he or she performs the desired movements.

Geert Caenen of KU Leuven talked about his work on making image-based structure-from-motion (SFM) modeling available to archaeologists and other producers of virtual heritage. (Slides: PDF). See Fig. 8. The work has involved a considerable systems engineering effort to make the SFM pipeline robust and usable for non-computer vision people. The basic steps are determining an ordering for the images, computing coarse scene geometry and camera positions using SFM, then refining the geometry by revisiting the images at higher resolution with a dense stereo algorithm. In addition, the SFM part has been integrated with 3D geometry processing for mesh merging and global model building developed by other research groups. The current system offers a web based interface that allows users to submit (unordered) sets of images of a scene. Each data set is processed as a background job distributed over a large set of KU-Leuven workstations. When processing has been completed the user is notified by email and can download the model if processing was successful. If no model could be generated the system sends an error message describing the point at which processing failed. A model viewer was also developed as a part of the project. Several results were shown, including one that captured a Dutch barn by parts and merged the results into a unified model.

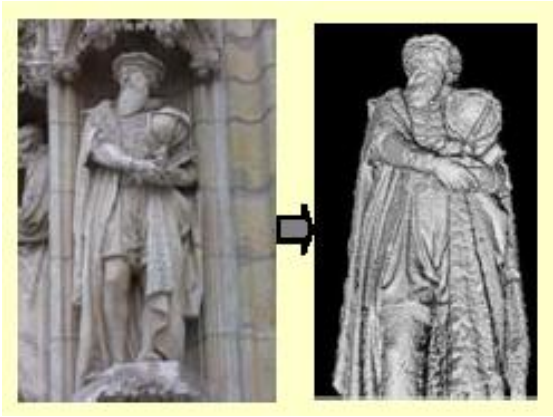

Figure 8: Several photos (left) are used to build a textured 3D digital model using SFM.

2.6 BIRS-Banff centre hands-on interaction

The above interaction sessions were followed by joint BIRS and Banff Arts Centre demo, lab and poster sessions. In the demo sessions the speakers showed how to apply their methods and algorithms to real world imaging data. Fig. 9 shows an example of the interaction. The interaction session turned out to be an excellent forum for further inquiry and discussions, where participants automatically grouped themselves informally into small discussion groups.

Another goal of the interaction sessions was to provide 3D models of real world objects and people for the Banff New Media Institute modelers and animators. Several of them came to the lab sessions and brought objects to be captured. We provided lab setups for geometry and appearance capture at several scales, ranging from small objects to people and people sized objects – see Figs. 10 and 11.

In another session we were able to tour the Banff New Media Institute and see demos in its audio, video and 3D studios. The modelers and animators showed us current work both related to art, such as some 3D cave visualizations done for a New Media Artist, and education such as a quantum physics game illustrating secure quantum communication.

Figure 9: Geert Caenen, demonstrates the KU Leuven 3D reconstruction software using SFM and dense stereo to a group of participants.

2.7 Variational Methods

Jan Erik Solem of Malmö University presented a general framework for variational formulations and level set methods in computer vision problems. (Slides: PDF). The framework consists of three steps: (i) selecting an appropriate energy functional, (ii) finding an initial starting point, and (iii) choosing a specific optimization method and stopping criterion. The level set method was used as a standard tool for solving the resulting energy minimization problems. A number of applications to both curve and surface problems were presented.

One frequently occurring example is the problem of recovering 3D models of a scene given only a sequence of images. Standard level set methods only allow closed surfaces to be reconstructed. By combining two different level sets, one describing the surface and the other used as a cut-off surface, open surfaces can also be treated. The standard optimization procedure is to calculate the Euler-Lagrange equations of the energy functional and use them as the right hand side in a system of partial differential equations. It was shown that this method can indeed be interpreted as a gradient descent procedure once the gradient of a functional is properly defined.

Yuri Boykov of the University of Western Ontario presented a framework for unifying continuous methods such as level sets and discrete methods such as graph cuts. (Slides: PDF, PPT). Among a multitude of approaches, the level set and graph minimal cut methods have emerged as two powerful paradigms for computing image segmentations. These methods are based on fundamentally different image representations. Level sets are formulated as infinite-dimensional optimization problems on a spatially continuous image domain, whereas graph cuts are defined as minimal cuts of a discrete graph representing the pixels of the image. Yuri showed that graph cuts can be an efficient tool for the local or global optimization of several computer vision functionals for geometric surfaces that are currently addressed mainly with variational methods based on gradient flow PDEs. He also showed how to use the Cauchy-Crofton formula to construct a grid-graph to approximate any given Riemannian metric up to a desired accuracy. Finally, he presented an integral approach to gradient flow where the max-flow algorithm is used to construct an optimal step of a fixed length.

Todd Zickler of Harvard University presented a method based on appearance decomposition for image-based shape recovery. (Slides: PDF, PPT). Image-based reconstruction systems are

Figure 10: Left: A participant, Peter, being captured as a 3D model. Right: Neil Birkbeck reconstructs his 3D geometry and dynamic texture using the U. of Alberta capture system.

Figure 11: The computed 3D wire-frame model and some textured renderings.

designed to accurately recover the three-dimensional shape of a scene from its two-dimensional images. The reconstruction problem is ill-posed because images do not generally provide direct access to 3D shape. Instead, shape information is coupled with additional factors such as illumination, pose and surface reflectance. Shape recovery typically requires assumptions about reflectance, for example that surfaces are Lambertian. When such assumptions are violated the accuracy of the recovered shape can be compromised.

Todd Zickler presented two techniques for recovering shape that relax the assumptions about surface reflectance. Both are based on the notion of an 'appearance decomposition'. By decoupling some of the factors that determine an image (shape, reflectance, illumination and pose), he obtained more direct access to shape information, greatly simplifying the reconstruction problem. The first part treated Helmholtz stereopsis with a focus on recent calibration work to make this reconstruction technique more practical. The second part presented a family of color-based photometric invariants that extend the applicability of Lambertian-based reconstruction techniques (structure from motion, stereo, photometric stereo, shape-from-shading, etc.) to a broad class of specular, non-Lambertian scenes.

Figure 12: Our visit to the Banff New Media Institute.

2.8 Scene Reconstruction II

Geert Caenen of KU Leuven presented a general framework for using generative image models in computer vision applications. (Slides: PDF, PPT). Generative models have already shown their value in computer vision. They explicitly model the image formation process in terms of different imaging parameters and the unknown scene, allowing a number of image-based inference problems to be solved by inverting the process. The probabilistic nature of the framework allows for the introduction of prior assumptions that can express coherence of the data, outliers and much more.

Geert Caenen introduced the basic mathematical building blocks and tools for solving the aforementioned problem (notably the E-M algorithm). He then demonstrated the genericity and modularity of the framework by discussing various applications including image registration, depth computation and 3D-reconstruction, illustrating these with practical examples.

Stefan Roth of Brown University presented the novel concept of specular flow and applied it to surface structure recovery. (Slides: PDF). In scenes containing specular objects, the image motion observed by a moving camera is an intermixture of the optical flows resulting from diffuse reflectance (diffuse flow) and from specular reflections (specular flow). Stefan Roth formalized the notion of specular flow with a few assumptions, showed how it relates to the 3D structure of the world, and developed an algorithm for estimating scene structure from 2D image motion.

Unlike previous work on isolated specular highlights, he used two image frames and estimated the semi-dense flow arising from specular reflections of textured scenes. A parametric model was used for the image motion of a quadratic surface patch viewed from a moving camera. The flow was modeled as a probabilistic mixture of diffuse and specular components and the 3D shape was recovered using Expectation-Maximization. Rather than treating specular reflections as noise to be removed or ignored, it was shown that the specular flow provides additional constraints on scene geometry that improve the estimation of 3D structure when compared with reconstruction from diffuse flow alone. The method was illustrated on a set of synthetic and real sequences of mixed specular-diffuse objects.

2.9 Scene Reconstruction III

Yuri Boykov of University of Western Ontario presented a photoflux functional for image segmentation. (Slides: PDF, PPT). Recent advances in image segmentation have shown that using flux based functionals can significantly improve the alignment of object boundaries. Yuri Boykov proposed a novel *photoflux* functional for multi-view 3D reconstruction that is closely related to the properties of photohulls. Since the photohull prior can be combined with regularization, the work unifies two major groups of multiview stereo techniques: *space carving* and *deformable models*. It retains benefits from both groups, allowing fine shape details to be recovered without over-smoothing while robustly handling noise. Photoflux provides a data-driven ballooning force that helps to segment thin structures or holes. Yuri Boykov proposed a number of different versions of photoflux based on global, local, and non-deterministic visibility models. Some forms of photoflux can easily be incorporated into standard regularization techniques, while new optimization methods were proposed for others. It was also shown that photoflux-maximizing shapes can be viewed as regularized Laplacian zero-crossings.

Patrick Hébert of the University of Laval presented a unified surface representation for 3D imaging and modeling. (Slides: PDF, PPT). Given a set of input data, the method builds a 3D model in the form of a geometric representation and an associated appearance model for the surface of each object. The process requires several steps that vary depending on whether the input in in the form of color images or range data. Patrick Hérbert considered model building as a unified process as opposed to a cascaded series of independent steps.

First he described an approach that was developed for the interactive modeling of surface geometry from range data. It involves several steps including acquisition, alignment, fusion, surface reconstruction, visualization and compression. A necessary condition for interactive modeling is that the computational complexity of each step should be linear in the amount of data acquired. The gradient of the signed distance field can be recovered directly from the range data and each subsequent step benefits from this representation.

Secondly, he described a new approach for modeling surface appearance. When the aim is to produce a model for visualization purposes alone, it is not essential to recover accurate geometry so long as the appearance remains photorealistic. Instead of assuming a specific reflectance model, he adopted an image-based approach that uses data acquired from a camera that is moved around an object to produce a light field from a large set of calibrated images. He proposed a frequency-based criterion to estimate a light field parametrization surface that is well adapted to the object and the set of views.

Li Zhang of Columbia University presented a space-time approach to 3D photography. (Slides: PPT). Recovering the 3D structure of a scene from photographs is an important problem in several areas, including computer vision, computer graphics, and robotics. Two fundamental challenges in 3D photography are the accurate reconstruction of scenes with complex occlusions and of scenes containing dynamic objects. Li Zhang presented a space-time approach to these two problems that exploits the temporal variation of spatial visual cues such as defocus and stereo.

He first presented a temporal defocus method that reliably recovers the 3D structure of a scene, regardless of its occlusion complexity. Then he presented a space-time stereo method that accurately reconstructs objects that are deforming over time. Both methods significantly outperform the state-of-the-art techniques for 3D sensing. Finally, he demonstrated several applications of the proposed methods to computer graphics, including image refocusing, video composition, expression synthesis, and facial animation

2.10 Visual Recognition and Learning

David Nistér of the University of Kentucky presented an impressive local appearance based image indexing scheme that scales efficiently to very large image databases. (Slides: PDF, PPT). The method reliably recognizes music CD covers from a 50 000 CD database in real time on a portable computer. A larger 6 processor desktop system can search a 10^8 image database in less than 6 seconds. The method is also being used for image matching in a city-scale visual reconstruction system. The scheme uses affine invariant local image descriptors for robustness to background clutter, occlusions and changes of view point, indexing these high-dimensional descriptors using a novel hierarchical vector quantization tree to provide rapid calculation and fine visual discrimination. The use of high-dimensional descriptors (128-D), a fine subdivision of descriptor space (10^6 or more leaves), a modest branching factor each node (10-16× – neither binary nor large), multi-level voting and careful vote weighting together lead to a dramatic improvement in retrieval quality.

More generally, many media processing applications would benefit from data structures that can efficiently index large sets of uncertain high-dimensional descriptors, but this appears to be a difficult research problem owing to the curse of dimensionality. A large number of structures including various kinds of spatial subdivision trees (kd-trees, quadtrees...), mixtures of trees and locality sensitive hashing have already been tried in this context with mixed results at best, but the hierarchical vector quantization scheme proposed here appears to offer hope for progress.

Vincent Lepetit of EPF Lausanne described a visual tracking method based on the visual recognition of characteristic local image regions ("keypoints"). (Slides: PDF). The method is trained using a number of views of the target object, after which it tracks the object's pose in real time. In both phases distinctive keypoints are detected in the image at multiple scales and image patches around them are extracted. The training method incrementally learns one visual class for each observed keypoint, so that the keypoints seen during tracking can be classified to the correct class (matched to the right source keypoint) despite changes in viewpoint and lighting. The keypoint correspondences then allow the object pose to be recovered. A novel classifier is used: a forest of randomized decision trees with internal decisions based on elementary pixel comparisons and probabilistic merging of tree outputs. This method combines rapid learning and execution and on-the-fly addition of classes with reliable classification over a large set of keypoint classes.

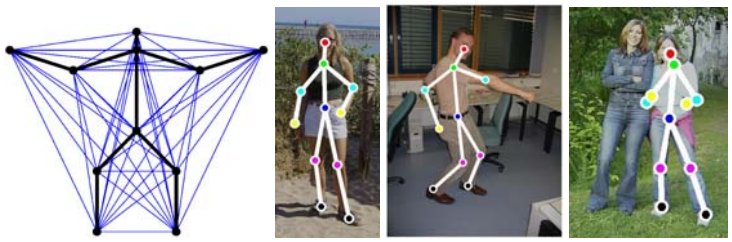

Figure 13: A graphical model showing the interactions between body members that are taken into account, and several examples of recovered poses, for Martin Bergtholdt's method.

Martin Bergtholdt of Mannheim University presented his work on learning probabilistic graphical models for recognizing classes of objects with highly variable geometry and appearance such as humans. (Slides: PDF, PPT). The method uses a parts-based representation of the class, learning both a statistical appearance model based on a local detector for each part

and pairwise relationships between parts. The problem of detecting instances of the object in the image then becomes one of finding optimal assignments of parts to image locations given the complete graphical model. This inference is done either exactly with an A^* search based on some interesting new admissibility heuristics, or (for larger networks) approximately using Belief Propagation. Experiments on face detection and on human detection despite complicated articulated body poses show the promise of the approach, which is illustrated in Fig. 13.

Greg Mori of Simon Fraser University presented his work on estimating unusual human poses from single images. (Slides: PPT). The main approach discussed begins by grouping the image pixels into 'superpixels' – small homogeneous regions that almost surely lie on just one object – as a means to reduce the computational complexity of later steps. It then uses learned part detectors to assemble these into image segments that represent possible positions for limb and body members, after which an efficient combinatorial search is applied to find the best assembly of the detected segments into a coherent human body. For an example see Fig. 14. Greg also described a variant based on combinatorially optimizing the image placement of an explicit 2D body model under the assumption that its joint centres lie at superpixel centres.

Figure 14: The superpixel-assembly based method for human pose estimation from a single image developed by Greg Mori's group. The images show respectively the input image, its segmentation into superpixels, a possible limb detection and the human body segments that were finally recovered.

Phil Torr of Oxford Brookes University presented his group's work on solving Markov Random Fields (MRF's) using Dynamic Graph Cuts and Second Order Cone Programming (SOCP) relaxations. (Slides: PDF, PPT). MRF's are probabilistic graphical models representing networks of individually simple but interconnected variables. They are used for many purposes in vision including image segmentation, stereo and model-image matching and assignment problems. The high dimensionalities of their state spaces make many calculations difficult, one of which is finding the optimal (Maximum Likelihood or Maximum A Posteriori) state assignment to the variables. The talk presented two methods for finding optimal or near-optimal assignments in particular cases.

For many practical MRF's with binary variables, the optimization can be reformulated as a graph cut problem and solved using flow based algorithms such as augmenting paths, however despite intensive work this computation remains relatively costly. Often we need to solve a series of similar problems, for example when processing videos or because non-binary MRF's can be approximately solved using a series of binary problems or because different parameter values need to be tested. The first part of the talk presented Dynamic Graph Cuts, a method for solving a similar problem by reusing the search trees computed for the augmenting paths in the current problem. The method often saves a great deal of computation, especially in applications such as video segmentation where many of the problems in the sequence are very similar.

The second part of the talk presented a generic method for approximating MAP states in MRF's by reformulating the problem as a 0-1 integer quadratic program (QP) and solving

this using relaxation. Initial work used a semidefinite programming relaxation, but a newer second order cone programming relaxation scheme gives similar or better results in much less time. Unlike other popular schemes such as Belief Propagation or Graph Cuts, convergence is achieved in just a few iterations. The method was illustrated on subgraph matching for object detection in triangulated images and on pictorial structure matching (probabilistic networks of parts with uncertain relative positions).

Jim Rehg of Georgia Tech presented his group's work on learning optimal rejection cascades for object detection. (Slides: PDF, PPT). Rejection based methods have shown themselves to be one of the most successful and efficient approaches to object detection. They sweep the image with a detection window, trying to save computation by rejecting windows that definitely do *not* contain the object as quickly as possible. For this they use a chain or tree of classifiers, each trained to handle the (increasingly more difficult) cases left undecided by the previous level. Each stage depends on the previous ones and classifier training is a relatively complex and expensive process, so optimizing such chains to meet a given set of final performance criteria is not simple. The talk described a probabilistic look-ahead predictor for the performance of the full rejection chain given the performance of the currently-trained elements of it. The method allows the algorithm to make local decisions about classifier quality and threshold during training that lead to good overall detector performance.

M. Alex O. Vasilescu of MIT Media Lab gave a talk about tensor (multilinear) extensions of Principal Components Analysis (PCA) and Independent Components Analysis (ICA), two traditional matrix-based linear dimensionality reduction methods that are often used in vision and graphics. The methods apply to data sets that have several dimensions of variation, for example sets of images of human faces with varying pose, lighting, expression and identity.

The tensor analogue of PCA is based on the "M-mode SVD" decomposition algorithm, which essentially uses a series of Singular Value Decompositions (SVD's) to apply independent rotations to each axis of variation of the data tensor, moving as much of the "energy" (squared norm) as possible into the first few components along each axis. Unlike the single axis of variation case (conventional PCA using SVD), the resulting reduced "core tensor" is not usually diagonal, although it often exhibits rapid decay along many or all of the axes. This makes truncated forms of it useful for data approximation. Projecting an incoming data vector (here an image) onto the core tensor axes allows the effects of the different influences to be separated to some extent, for example removing pose and lighting variations to get a more canonical image encoding intrinsic facial appearance (identity). Similarly, Multilinear Independent Components Analysis (MICA) generalizes conventional linear ICA by applying transformations to each axis that are designed to detect and emphasize non-Gaussian behavior such as long tails.

The methods were demonstrated in the context of facial image biometrics under changing facial geometry, expression, lighting and viewpoint. In this application, the "TensorFace" (M-mode PCA) and "Independent TensorFace" (MICA) representations provide significantly improved recognition rates relative to standard PCA and ICA – see Fig. 15. A second demonstration, TensorTextures, described an image based rendering technique that learns a multilinear generative model for surface appearance from a sparse set of example images. The model captures interactions between viewpoint, illumination and geometry including complex details such as self-occlusion and self shadowing. A third demonstration extracted human motion "signatures" useful for motion recognition and in computer graphics character animation from human motion capture data.

2.11 Human Motion

Cristian Sminchisescu of the Toyota Research Institute Chicago gave a talk on Bayesian inference algorithms for estimating 3D articular human motion from monocular video sequences. (Slides: PDF). The problem is difficult because the human body has many degrees of freedom

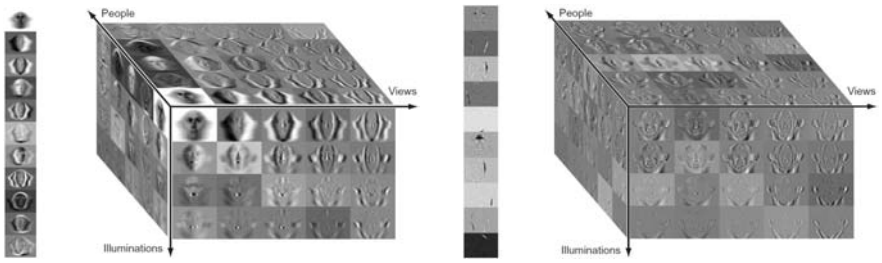

Figure 15: M-mode tensor representations of a face data set with identity, lighting and facial pose variations. The left panel shows M-mode PCA, the right one M-mode ICA. On the left of each panel, the corresponding non-tensor representation (i.e. normal PCA, ICA) is shown for comparison.

and these are difficult to observe in monocular images owing to occlusions and depth ambiguities. Body tracking research has traditionally used generative (forwards) models that predict image observations based on a 3D body model but this rapidly runs into problems of ambiguities and local minima because the inverse problem is inherently multi-valued. Recent work has inverted this, directly learning to predict 3D body pose from image observations (discriminative or inverse modeling), typically regularized by some form of prior on typical human poses or motions to stabilize the solution and reduce the degree of multi-valuedness. The focus of the talk was a model of this kind called BM^{3}E based on embedding a multi-valued Bayesian mixture of expert inverse model in a Markov chain tracker. The talk also showed how to use kernel-based nonlinear dimensionality reduction to reduce the state space to be estimated to a more manageable dimension, and how to jointly learn discriminative and generative models to provide more resistance to tracking failures. Some examples of static poses recovered by the method are shown in Fig. 16.

Figure 16: Some human poses recovered from single images by Cristian Sminchisescu's mixture of experts method.

Raquel Urtasun of MIT described how to use Gaussian Processes[2] (GP's) to learn prior models of human pose and motion for 3D person tracking. A Gaussian Process Latent variable Model (GPLVM) provides a low-dimensional embedding of the human pose and defines a density function that gives higher probability to poses close to the training data, while a Gaussian Process Dynamical Model (GPDM) uses a second GP to provide a complex dynamical model.

[2]GP's are infinite dimensional probability models taking the form of Gaussians in some linear function space. Practical applications involve conditioning on a finite number of training observations to obtain a posterior model that then allows point-predictions to be made using standard matrix calculus. GP based regression is computationally expensive but it provides good levels of generalization from small training sets and useful model uncertainty estimates, both of which are very useful in the high-dimensional limited-data learning problems considered here.

Bayesian model averaging allows both the GPLVM and the GPDM to be learned from relatively small amounts of training data, and they provide graceful generalization to motions not seen in the training set. Tracking is then formulated using MAP estimation on short sequences of poses within a sliding temporal window. These priors allow effective tracking of a range of human walking styles, despite weak and noisy image measurements and a very simple image likelihood. Fig. 17 illustrates the process.

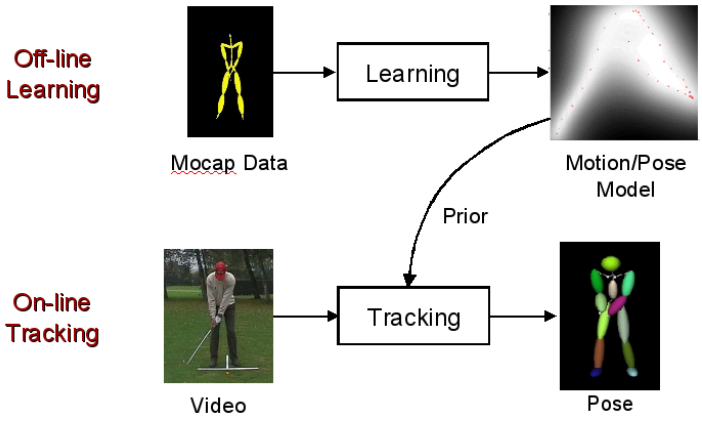

Figure 17: The stages of Raquel Urtasun's Gaussian Process Dynamical Model based human tracking framework.

David Fleet of the University of Toronto described his group's work-in-progress on physics based priors for modeling walking dynamics for 3D human tracking. (Slides: PDF). The current motion priors for tracking are typically simple kinematic models that can not handle issues of balance and contact dynamics. The new method uses a dynamical model inspired by the passive anthropometric walker of (McGeer 1990) and (Kuo 2001,2002) – a simple 2D two phase analytical model including toe-off impulses and a torsional spring in the hip, that manages to simulate many aspects of human walking quite well even though it has no knees. The new model (which does have knees) is used to track the lower bodies of walking people from monocular video sequences, using an on-line sequential Monte Carlo tracking procedure to infer kinematic, dynamic and anthropometric state variables. The tracker tolerates significant occlusions and handles people walking straight and turning.

3 Summary and Prospects

Overall, we had very positive feedback about the workshop. The presentations were universally well received and the ensuing discussions were often very lively. The Interaction with the Banff Centre New Media Arts people was a new and interesting experience for both parties. The friendly atmosphere of the facilities and the excursions of the group to the surroundings of Banff stimulated informal discussions and thus contributed to the success of the meeting. A number of new collaborations appear to have been started as a result of these interactions and many of the participants have asked us whether there are any plans for a follow-up event.

It is clear that computer vision is a field that is currently making rapid progress, in part owing to the widespread adoption of advanced mathematical techniques including partial differential equation and random field models, mathematical programming, statistical learning,

tensor decomposition, dimensionality reduction, Gaussian processes and even Galois theory. Our ability to reconstruct detailed models of static and dynamic scenes from images and to recognize individual objects, object classes and human motions from images has increased dramatically over the past decade and this trend is expected to continue for at least the next one. Although the problems studied in computer vision are seldom very pure from a mathematical point of view, they have great diversity and many mathematically rich aspects, so further opportunities for interaction between the two fields would be highly desirable.

We see a number of emerging topics that are likely to provide especially rich areas for interactions over the next few years:

- Advances in the various different ways of attacking the minimization of *large-scale energy models* (mathematical programming, graph cuts, PDE and level set methods...), and more generally in statistical computations on large-scale graphical models, will lead to improved scene reconstruction and motion estimation methods, and also to improved model-image matching and object recognition methods.

- The *interaction between computer vision and computer graphics* will continue to increase, especially in areas where images are used as sources for rendering (capture of detailed surface geometry and reflectance properties; modeling of large scale scenes and natural phenomena; light based models such as lumigraphs; capture of human appearance, expression and movement). As the scale, degree of realism and mathematical sophistication of such models continues to grow, it will become ever more necessary to involve mathematicians in this collaboration.

- More realistic *statistical models of natural images* will lead to more reliable and invariant image features and a better understanding of what is needed for reliable visual recognition. This is an extremely complex high-dimensional statistical modeling problem.

- The basic problem of *representing 2D and 3D form* has been one of the mainstays of vision research for decades, but is not yet fully resolved. In one research direction, variational methods are combined with increasingly precise physical models. In another, methods and results from geometric computer vision are revisited and refined into new methods.

- In terms of impact on society, computer vision is poised to follow its sister disciplines image processing and computer graphics into mainstream use. Most notably, this means that new classes of users will be introduced to computer vision techniques. Today just about any digital camera owner can apply advanced image processing techniques, e.g. using PhotoShop, and groups of artists without science degrees can model and animate computer graphics worlds. A challenge for computer vision is not only to develop methods and algorithms, but also to make them usable to the general public, for example making it possible for a layperson to compute a 3D model from photos.

Finally, in the name of all of the organizers and participants, we would like to thank PIMS, BIRS and their sponsors for giving us the opportunity to organize this event in such prospicious surroundings.

Figure 18: Some of the participants on a day trip at Moraine Lake.