i3DPost - Second Annual Report, 2008-9

www.i3dpost.eu
The i3DPost Project is formed of six companies and educational institutions that are Europe's leading specialists in their fields. The Foundry Visionmongers ltd., Centre of Research and Technology Hellas, Informatics and Telematics Institute (CERTH-ITI), BUF Compagnie, University of Surrey (UniS), Quantic Dream (QD), The University of Dublin (TCD).

i3DPost will develop new methods and intelligent technologies for the extraction of structured 3D content models from video, at a level of quality suitable for use in digital cinema and interactive games.

The research will enable the increasingly automatic manipulation and re-use of characters, with changes of viewpoint and lighting.

i3DPost will combine advances in 3D data capture, 3D motion estimation, post-production tools and media semantics. The result will be film quality 3D content in a structured form, with semantic tagging, which can be manipulated in a
Summary of Activities

Key scientific & technical innovations

The innovation that i3DPost proposes will produce a step change in the quality of 3D video and postproduction, combined with semantic metadata, to achieve practical implementations of intelligent media technology at a quality level that meets the requirements of the film, games and professional media industry.

The goal is to create methods that make it possible to edit and alter filmed sets and actors – their faces, clothes and performances – as if they were animations, to allow their simple manipulation and reuse within post-production and the retargeting of filmic content to interactive platforms. The integration of semantic technologies with next-generation approaches to 3D video capture and postproduction offers a means of changing the way the media industries work, and opens the way to the repurposing of film and video material to create new programme experiences.

The result will be film quality 3D content in a structured form, with semantic tagging, which can be manipulated in a graphic production pipeline and reused across different media platforms including films (either 2D or 3D), computer games, music and promotional videos.
The overall objective is therefore to integrate 3D information extracted from the visual scene into all stages of the postproduction pipeline, supported by semantic metadata, at a quality level that meets the requirements of the film, games and professional media industry. This will be achieved by the invention of methods for recovering intelligent, structured content from on-set filming, and incorporating them in capture environments and software tools that enable the simple manipulation and reuse within post-production and the retargeting of content to interactive platforms.

Having completed the Initial set-up phase, the establishing of the scenarios of use, proof of the concepts in the main areas of research, and development Proposals, the work on the i3DPost Project has continued with evaluation and towards the completion of the Initial Protoypes.

Portable Multiple Camera Capture System

The University of Surrey has developed a portable system for multiple view video capture on location, based on 8 HD Canon Camcorders. This system allows synchronised recording of 1080p video in a compressed format to tape or with uncompressed HD-SDI output to an external disk. Camera timecodes are synchronised and locked in an initial setup phase. Recording can then be performed independently for each camera using the internal timecode for synchronisation. Alternatively recording for all 8 cameras can be controlled from a laptop using the firewire interface. Initial test sequences for indoor and outdoor scenes have been successfully recorded validating the camera synchronisation and portable operation.
Spheron HDR Camera

Spherical stereo disparity maps for indoor and outdoor scenes
Reconstruction of an outdoor Cathedral scene
3D Feature Analysis & Understanding

The work in this area deals with developing algorithms that unlock the correlation between related views of the same scene to facilitate both 3D modelling and 2D motion estimation.

**Multiview fusion for 3D point cloud estimation**

The team at Trinity College has been focusing their research on exploring various optimisation strategies for depth map extraction. Multiple techniques have been implemented and made available to the partners in the form of functions callable from MATLAB.

A new technique to refine depth maps based on Graph-Cuts has been developed. It allows the fusion of multiple candidate depth maps at once which reduces the computational complexity of depth estimation. Restricting the computational load is essential when working with High Definition images. This research has been accepted in the peer reviewed CVMP'09 conference.

Another aspect of their work has been to bridge the gap between sparse matches obtained from keypoint matching and dense disparity/motion estimation estimation. A paper based
This year also saw the start of research on incorporating Geodesic Distances into spatial inference. This has led to developments in sparse to fine segmentation, disparity and motion estimation. In addition, they have developed a new tool for interactive segmentation based on this idea. This new technique combines colour and spatial information with user strokes to bring a measure of interactivity to matting in the post-production environment.

**Soft Geometrical Constraints in Image Centric Manipulation**

Trinity College has also provided an alpha matte estimation tool which contained the use of multiple views to guide the alpha matte estimation. They completed the alpha extraction tool by adding a Foreground/Background matte estimation, which uses the alpha matte to push the estimation of the background and foreground values in the areas where the blend occurs.

A new background subtraction algorithm has been developed. Background subtraction produces foreground objects mattes which are then use to extract visual hulls from the object. This new algorithm is an intuitive extension of the work by Levin et al. on Matting which has been used in the past. Initial results suggested that significant quality improvements can be gained.

TCD has packaged two of the matting algorithms it has developed into NUKE plugins. They now enter the testing phase in the quarter.

Work on automated object segmentation based on long stable feature tracks has been ongoing throughout the second year. An early part of that work has been accepted in the SPIE VCIP onference in 2010. This work uses segmentation of sparse motion trajectories to yield dense segmentation of pixels throughout the sequence. An interactive version of this work will also be packaged into a NUKE plug-in in due course.

**The Foundry** has continued to improve the software components for camera calibration, camera pose estimation, and unstructured 3d points.
The framework can now solve for multiple-view sequences. The narrow-baseline solves are reliable enough for production use for feature-film stereo-camera captures. Further algorithmic work has been carried out on developing wide-baseline matching of SIFT features to calibrate a multi-camera setup directly from the input images and to generate semi-dense / dense point clouds. As proof of concept this has been tested using the studio capture data-set provided by the University of Surrey with a known ground-truth camera calibration.

Prototypes have also been produced to show the insertion of semantics information – currently bounding boxes on objects of interest – into the point cloud.

**Set, Actor & Face Modelling**

The purpose of this area of research is to generate high quality representations of sets and actors from the captured data taking into account the dynamics of skin and clothing.
**Set Modelling** by The University of Surrey

Set reconstruction from spherical stereo image pairs has been developed giving a dense 3D scene reconstruction with per-pixel depth and colour information. Spherical image capture is performed at a resolution 10000x5000 using the Sphereon VR high-dynamic range line scan camera acquiring a vertically displaced stereo pair. Automatic stereo matching is then performed to estimate the depth for each pixel. Subsequent processing simplifies the raw stereo reconstruction to produce an efficient scene representation with high-resolution images texture mapped onto the model to reproduce surface appearance detail. The approach has been used to reconstruct indoor and outdoor scenes.

**Actor Modelling** by The University of Surrey

Research on actor modelling is developing methods to transform raw 3D reconstructions of actor performance acquired using multiple view video into structured temporally coherent mesh sequences representations. Temporal alignment is performed by tracking visual features on the person and using a constrained mesh deformation to map a single consistent mesh onto each frame of the sequences. A volumetric Laplacian mesh deformation technique is used to transform the initial mesh based on a sparse set of visual features. The approach also allows the estimation of joint positions based on initial alignment of a skeletal joint structure with the initial mesh.

**Face Modelling** by The Foundry

Dense disparity estimation for narrow-baseline video sequences has been extended to relatively wide-baseline multi-view camera configurations. Wide-baseline matching of SIFT features from research work on 3D Feature Analysis has generated camera-to-camera correspondences to build up the facial surface. The results are significantly more robust, which creates better facial
Multi-view camera calibration and point cloud estimation for the face has been combined with KLT feature tracking to create 3D surface trajectories. This has been used to prototype a virtual marker tracking system for facial animation. A set of "virtual markers" are defined in the video sequence and tracked to create a moving 3D point for each marker.
One of the aims of the project is to provide a semantically rich representation of multi-view videos and 3D scenes. This will be achieved by labeling the scene elements, i.e., identifying actors and important objects (e.g. props) and tagging them with actor identity (e.g. John Smith) and object class (e.g. chair), respectively. The semantic description will be further enriched with information regarding the actors’ facial expressions and performed actions or interactions. The resulting scene annotation will be used, in a proper metadata format, for the intelligent manipulation and organization of the content. Within this framework, a number of research directions have been followed during the second year of the project, most of them as a continuation of research that was conducted during the first year. In addition to the activities described below, a considerable amount of effort has been spent on thoroughly testing all methods developed so far on data recorded within the i3DPost and fine-tuning them accordingly.

**Human Activity Recognition**

The study of methods that use multi-view human activity videos so as to achieve view-invariant activity recognition was completed. The most important advantage of this approach is that activities can be recognised regardless of the movement direction of the subject.

*Figure 1 Frames corresponding to the same time instance, from a multi-view video sequence depicting a walking person.*
The methods are applied on the binary masks resulting from the coarse segmentation of the body and utilize fuzzy c-means clustering, linear discriminant analysis and fuzzy vector quantization. The two approaches that were devised in the previous year were thoroughly tested during the second year on data recorded within the project using eight synchronized high-resolution cameras. These data depict five movements (walk, run, bend, jump in place, jump forward) performed by eight persons. The approach that solves the view correspondence problem (i.e., the identification of the view angle, with respect to the body, for each camera) using a circular correlation based view rearrangement of the multiple views (so that they are consistent with the arrangement of views in the training data) achieved a correct classification rate of 78% whereas the method that tackles the view correspondence problem using a DFT representation achieved a much better rate of 90%. Furthermore research on the extension of the video-based algorithms that have been developed so far to 3D data has started. The input to these algorithms will be binary, voxel-based representations of frame sequences depicting the activity.

Figure 2 3D voxel-based instances (frames) of two sequences depicting human activities (bend and run).

In addition, a multiple viewpoint human activity video database was created. This database is to be used for the experimental evaluation of activity recognition and motion analysis algorithms and was recorded using four synchronized medium resolution color digital video cameras placed at the four sides of a parallelepiped. The recordings include 13 actions or combinations of actions performed by single subjects and 6 interactions performed by pairs of subjects, over fairly uniform background. The database contains 39 subjects (64% men and 36% women), 30 of which repeated a small subset of the above actions after two to three days in order to provide variance for
biometric purposes. The database is currently being post-processed and will subsequently be released in the public domain.

**Person Identification Based on Movements**

Work has been performed on person identification using movements other than walk or run such as jump, bend etc. This is a highly novel approach because, despite the fact that gait-based biometrics have attracted a lot of attention, most researchers focus on walking or running. The algorithm that was devised comprises of two parts. The first part consists of a classifier that assigns each video to one of the R studied movements. The second part is a set of R classifiers (one for each movement) that assign a video of the corresponding movement type to one of the persons that have been used for the training. For all classifiers, the Fuzzy Vector Quantization / Linear Discriminant Analysis algorithm that was developed for single-view movement recognition within the project has been used, each time trained with different training sets and labels. For a video depicting an unknown person that performs an unknown movement, the movement classifier is used to classify the video to one of the movement types (walk, run, etc) and then the video is forwarded to the respective movement-specific person classifier that decides for the person that performs the movement. In case of videos depicting multiple movements of the same person, the results from the R classifiers are fused. Experiments were performed on a database that contains videos of 9 persons performing 7 movements (walk, run, skip, gallop sideways, jump jack, jump forward and jump in place). The results show that movements such as skip or jump in place can provide very good person recognition results (93% and 92% respectively), even better than those obtained from walking. In addition, fusion of the results from different movements leads to almost perfect person recognition. To verify the results, testing on bigger databases should be performed.

Object Recognition

Two different object recognition methods were developed in the second year. The first uses images of an object taken from multiple cameras (currently two) with known relative positions. The method is based on vocabularies of local
features—it is essentially an extension of a bag-of-keypoints approach—, and thus robust to partial occlusion and many transformations. It was tested on a database of furniture images and was found to have superior performance to fusing the results of single-view methods applied separately on each view.

The second method is based on a single view of the object, is more discriminant but less robust and thus complements the above one. It is based on the segmentation of the object image into a number of discriminant regions which are different for each class, with a class being defined as a combination of object and viewpoint. The regions are chosen so that they best characterize an object’s class vis-à-vis all other classes. Again, promising results were achieved on a database of furniture images. Both methods are still under study.

Figure 3 Features used in the first object recognition method.

Frontal Face Detection

In the area of frontal facial view detection, we adapted the object recognition approach that is based on discriminant regions (see above) so that it can be used for the characterization of facial images as frontal or non-frontal. In this case, the corresponding classes that are involved are the frontal view, as well as a number of side views. The first results are very promising. The algorithm is currently tested in more challenging data. The method can be used as a pre-processing step to face and facial
expression recognition techniques that operate on frontal views. By using this method, the camera that captures the frontal view of a face or the frames of a video that display a frontal view will be detected and fed to a frontal face recognition or frontal facial expression recognition technique.

Facial Expression Recognition

A new algorithm for the recognition of the six basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) from single-view videos has been developed. The first step of the method is to track the grid nodes of the Candide facial model (superimposed on the face) from the neutral to the full facial expression in a video sequence.

Figure 4 Frames depicting neutral and full expression (happiness), with the Candide model superimposed on the face.

We then calculate the barycenter of the nodes and their principal components and map all nodes to the coordinate system defined by the barycenter and the principal components. Support Vector Machine classifiers are then used for recognizing the facial expressions. Due to the new coordinate system the method is robust to in-plane translation, rotation and scaling of the face. It also achieves very satisfactory results, being successful in approximately 89% of the tests in an industry standard benchmark database. Using the above method as a starting point, we also derived a method for facial expression recognition in 3D point cloud data. This method differs in that it does not operate on Candide nodes but on points contained in a cloud. Currently, the method is thoroughly tested.
We have also developed another novel method that operates on still images of the fully expressive face. The method is more accurate but also more fragile with respect to perturbations. Here, the 7-class (six basic expressions plus the neutral one) classification process is decomposed into all possible two-class (pair-wise) classification problems. Then, a unique set of features, based on Gabor filters, is extracted for each classification problem through an iterative feature selection process (that takes into account the effectiveness of sets of features in discriminating between the two classes) and Linear Discriminant Analysis (LDA) is applied to solve each two-class problem. The pair-wise classification results are then fused using a voting scheme in order to produce the final decision. In addition, an alternative decision fusion procedure that utilizes a Directed Acyclic Graph has been tested. The method was evaluated on an industry standard facial expression recognition benchmark database. The correct recognition rates for both variants of the algorithm are above other state-of-the-art methods applied on the same database.
Figure 6  Features related to happiness (b) and surprise (c) extracted from face images (a) using Gabor filters.

**Facial Feature Detection**

A facial feature detection algorithm was developed in order to provide landmark-based alignment of facial images for the face detection and facial expression recognition algorithms. The method is based on a novel split-and-merge image segmentation approach. As a result of this procedure, small areas resulting from the splitting step in the eyes and the mouth areas are merged together. However, due to differences in intensity and size, the eye and mouth areas are not merged with the facial area that surrounds them. This leads to the creation of at least three ellipsoidal regions in the face area. Geometric constraints are subsequently used to select the actual features of interest. Up to now the results are very promising.

**Intelligent tools for content manipulation**

Work continued through the second year to allow the project partners to use Nuke as the application platform for i3dpost technology. Nuke is The Foundry’s existing powerful compositing application that delivers high speed, an
efficient multi-channel scanline rendering engine, and a wide feature set. The camera calibration framework and the lens framework are now implemented in library modules with clear interfaces.

This application framework (Nuke) can also handle the passing of non-image data as meta-data through the processing tree, including suitable controls for the merging and overriding of meta data when combining multiple sources. The meta-data in the tree is available to external modules, whether implemented through the NDK or OFX interfaces to the framework.

Four prototype plugins have been developed for Nuke: Camera Calibration, Point Cloud Generation, Multi-View tracking & VirtualMarker. These prototypes have entered the evaluation stage with the User Partners. The feedback will define areas for further development.

Major Achievements

The technical scope of the project has been outlined above. The consortium has successfully completed it’s second stage of activities culminating in one commercially viable product on the market, and three additional plug-ins. ‘Ocula’ plugins are based on brand new disparity-mapping algorithms that have arisen from the research work carried out for i3DPost and allow artists to apply a multitude of adjustments to stereo image pairs. The Foundry won an iaward for ‘Ocula’ in November. The iawards are new and aim to become the BAFTAS of the Science and Technology world in the UK. Visit www.thefoundry.co.uk for further information.
Open-FX now has a proper independent legal framework. A committee is in place to review version 2.0 of the specification. The Foundry are pushing for adoption of the internal multi-view extensions as part of the standard for 2.0.

Promotion and Awareness

In addition to the user partners who are closely involved in i3DPost, an external user group with currently 10 members including post-production companies & games companies was formed and are regularly consulted.

I3DPost has been publicised at media events such as IBC, SIGGRAPH, NAB & CVMP and numerous conference papers and lectures have been given. (Please see website for details).

The website is frequently updated with project results and has an average of 10,000 hits per month, up from last years 7,000 hits per month.
Future Work & Exploitation Prospects

Based on the work done in the first two years, i3DPost will be refining the initial prototypes, evaluating the work and then working towards the final demonstrator.

Continued dissemination of the project results will be carried out at conferences and trade fairs as described in the Initial Operational Plan.

Further Information

Please visit the project website for up to date progress of our work.  
www.i3dpost.eu
Or contact the Project manager: Abi Bowman
abi@thefoundry.co.uk