Machine Vision and Applications (2003)
Digital Object Identifier (DOI) 10.1007/s00138-003-0123-4

Machine Vision and Applications

Computer vision for human modelling and analysis: position statement at HUMANS 2000

Adrian Hilton

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK

Abstract. This report summarises the author's views and experience on the application of computer vision technology for the modelling and analysis of people. The author conducted research which led to the first commercial booth system for capturing animated models of people for applications in games, multimedia and virtual reality. This research is ongoing with the aim of developing studio capture technology to enable photo-realistic capture of a person's shape, appearance and movement for broadcast production.

Key words: Human modelling – Visual reconstruction

1. Application requirements

Computer vision technology which can capture a person's shape, appearance and movement from a video image sequence has a large number of potential application domains. Examples of potential applications range from automatic body measurement for the clothing industry, through photo-realistic content production for film to human-computer interfaces to interpret expressions and gestures. Each of these application domains has specific requirements which include: accuracy, photo-realism, speed, interaction, reliability, cost and physical space. These criteria need to be considered in detail when developing a vision-based solution for a particular application. For example, in clothing applications, the dominant requirement is accurate body shape measurement, together with cost and space.

Application of vision technology in these domains requires consideration of the specific requirements of the particular application. This is critical as the limitations of computer vision are such that the solution must be tailored to address a specific problem. General solutions tend to provide results which fall short of the requirements of any particular application. For example, a general iterative shape reconstruction algorithm may be slow or not robust, whereas a domain-specific model-based approach may be considerably faster and more reliable but only applicable to the particular application. This does not prohibit research into generic methods for solving the open problems of computer vision. However, it does mean that in transferring this technology to address an application it is critical to consider the specific requirements together with the limitations of the generic vision algorithms in order to achieve a satisfactory system.

Principal requirements of particular domains in which vision-based human modelling and motion capture are applicable include:

Clothing: accurate shape measurement
Medical: accurate shape measurement, movement analysis and diagnosis
Film: photo-realistic virtual actors and markerless capture of natural movement
Games/location-based entertainment: low-cost realistic modelling
Human-computer interfaces: movement capture for gesture analysis

Computer-vision-based solutions are starting to be used in several of these application domains, most notably by the entertainment industry for both film and games where there is a strong demand for tools to aid in content production. In the remainder of this article, two such systems developed by the author and colleagues are discussed.

2. Modelling people for games and VR

An example application domain in which we have experience is the development of a capture technology suitable for modelling people in games and virtual reality applications [6,7]. The dominant requirements for this application are realistic appearance, low cost and automatic in a small capture space. The goal of our work was to develop a system suitable for use in an automatic booth system similar to a passport photo booth to generate 3D animated models of people which can be uploaded to the Internet for use in games, virtual reality and communication applications. For a mass-user application of this technology, cost dictates that the system should run as a stand-alone automatic system and be of a similar size to standard passport photo booths currently found in public spaces.

Correspondence to: A. Hilton (e-mail: a.hilton@surrey.ac.uk)

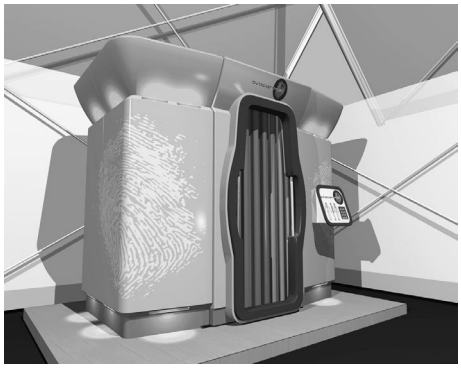
Prior to this work, commercial systems available for human modelling were focused on accurate measurement for use in clothing and large-scale anthropometric surveys. Manufacturers of body scanning system included Cyberware, Vitronic, Telmat, TCTI and Wicks&Wilson. These systems are highly expensive and do not provide an automatic method for generating animated models of people.

Due to the requirements of the games/VR application domain for automatic capture of realistic animated models, we developed a vision-based technology. This addresses the following application requirements:

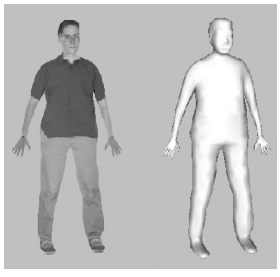
- Low cost
- Mass-user
- Automatic
- Fast
- Small size
- Recognisable models of people
- Animated
- Reliable operation for changes in size/clothing of person

The solution developed uses a model-based approach with a single camera. The subject is requested to stand in a fixed posture facing the camera and to the side. Even with this constraint the variation in pose between subjects is considerable. Silhouette images of the person are extracted and used to morph the shape of a generic humanoid model to that of a particular individual. To avoid problems of chroma-key, where the subject's clothing may be a similar colour to the background, a lighting panel is used to give a high foreground/background contrast, which allows reliable silhouette extraction irrespective of clothing. The morphed model is texture mapped from the front and side images to obtain a model with realistic appearance. Models can then be animated using the internal articulated joint structure of the generic model. Further details of this approach can be found in [7]. The result of this work is a system which achieves fully automatic generation of recognisable animated models of individual people suitable for games/VR applications. Due to the use of a single camera, the overall size required for the system is small. The use of prior models for human shape together with specific feature extraction algorithms for the known pose enable reliable reconstruction for large variations in size and shape. A closed-form analytic solution was developed for rapid reconstruction (under 3 s for the prototype system). These algorithms are specific to the particular application but achieve the specified requirements for a cost-effective solution.

This technology has been licensed to AvatarMe Ltd. (avatar-me.com) for the development of the first commercial booth system to capture models of people. Figure 1a shows the avatar booth system. A system has been installed at a location-based entertainment site in London throughout 2000 and used to capture a quarter of a million models of members of the general public. The captured avatars are shown on site in various computer-generated entertainment applications. Avatars can be downloaded from the Internet and used in applications for animated e-mail and popular games such as the 'The Sims' and 'Quake II'. Figure 1b shows an example textured avatar and the raw shape estimate from front and side view silhouettes. Figure 1c illustrates the use of an avatar in 'The Sims' computer game application converted by Vapour Technology Ltd.(vapourtech.com).



a Avatar-Me Booth (avatar-me.com)



b Example "avatar" of a person

c "The Sims" Games (Vapour Technology Ltd.)

Fig. 1. Generation of animated models of people "avatars" from photographs

This application demonstrates the need to tailor the use of computer vision to the requirements of a particular application. Through the development of model-based reconstruction from silhouette images of people with clothing, we were able to address the requirement for reliable automatic reconstruction for a large range of body size and clothing. Automatic reconstruction enables the booth to be used without an operator, which vastly reduces the running cost. The use of a generic model also provides the internal structure required for animation of the model, which is essential for the application domain. The use of a single-camera rather than multi-camera capture system allows a small size, which is a requirement for putting booths in a variety of public locations where space is at a premium. The result is a mass-user application of computer vision technology for capturing shape and appearance [7].

3. Modelling people for broadcast production

There is a strong demand in the film and broadcast industries for studio production technologies which enable video content to be more flexibly manipulated in post-production. Desirable post-production operations include manipulation of camera viewpoints or actor pose together with relighting of actors to match scene lighting, improved interaction of actors with virtual sets and adding virtual elements to real scenes. Recently computer vision research has been exploited to develop commercial tools for post-production in the film industry; examples include scene reconstruction from uncalibrated cameras for augmenting video of real scenes with virtual objects [14](2d3.com) and image mosaicing for removing objects from video of real scenes [9](imagineersystems.com).

The strict requirement for computer-vision-based techniques to be acceptable for broadcast content production is that the resulting content must be of video quality. Consumers of current broadcast content will expect the photo-realism of future productions to be at least as good as current programs. There is also a demand for new tools to produce highly realistic content for new media forms such as interactive entertainment and the Web. Computer vision techniques are required to reconstruct video-quality image-based or model-based representations of actor shape, appearance and movement. Requirements for actor content production in broadcast and film include:

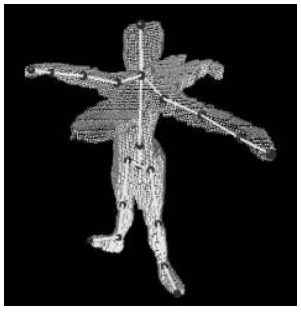
- Video quality photo-realism
- Simultaneous production of conventional 2D programs
- Markerless capture of shape and movement
- Realistic shape and movement
- Video-rate reconstruction for production feedback
- Synthesis of arbitrary viewpoints and trajectories
- Relighting of actor with virtual scene illumination
- Interaction of actors with virtual set and objects
- Simultaneous capture of multiple actors

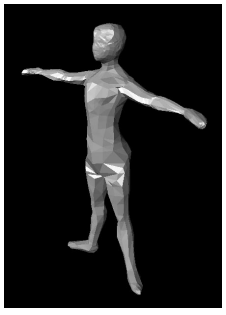
The requirements for video-quality and non-invasive capture are hard constraints on studio production tools, whereas other requirements are soft constraints. Conventional marker-based motion capture systems provide high-accuracy movement capture but do not allow photo-realistic production. The challenge for computer vision technology is to simultaneously reconstruct models of both the appearance and movement of actors either in a studio or, ideally, on location.

Current research, in collaboration with the BBC, is investigating the use of a multiple-camera studio to reconstruct photo-realistic actor models suitable for broadcast production [5,12,13]. This research has extended previous model-based techniques detailed in the previous section to a general N-view projective reconstruction framework. Again, prior knowledge in the form a generic humanoid model is used to constrain the reconstruction process and enable robust reconstruction in the presence of visual ambiguities. Figure 2a shows images of a dancer captured in a blue-screen studio using six broadcast-quality cameras. The number of camera views is limited due to camera equipment cost resulting in wide-angle views between cameras, which prohibits direct stereo correspondence using correlation. Standard blue-screen technology together with controlled illumination is used to separate the actor foreground from the background to sub-pixel accuracy. Foreground segmentation techniques for arbitrary backgrounds have been widely developed in computer vision. However, even with adaptive background models such techniques fail to reliably segment foreground objects with sub-pixel accuracy and are therefore not used for studio production. In this application, where visual quality is the primary metric, use of chroma-key technology to achieve the best results is appropriate.

Our research has extended the model-based vision methodology [12,13] to reconstruction from multi-view video of actor shape, appearance and movement. The model-based approach enables robust reconstruction of actor shape and appearance in the presence of visual ambiguities. The visual hull reconstructed from silhouette images of the human body reconstructed using a small number of camera views is ambiguous due to self-occlusion, as illustrated in Fig. 3b. The use of a model-based approach together with local shape constraints [13] which are invariant to scale enables robust reconstruction of an approximate shape model from the visual silhouettes. The initial model reconstruction is illustrated in Fig. 3c. Note the correct reconstruction of the dancer's chest and back despite the self-occlusion. This illustrates the advantage of model-based reconstruction over previous multi-camera volumetric approaches [8,10] where a large number of views were required to achieve correct reconstruction of human shape during movement. Model-based projective reconstruction techniques have previously been developed [1, 4] which also demonstrate the utility of strong prior shape models for reconstruction from a limited number of views in specific application domains.

a Multi-view images (three of six views)





b Visual hull. **c** Control model fit
d Animation in street scene

Fig. 2. Multi-camera studio reconstruction of dancer model

Given the initial approximation, model-based multi-view bundle adjustment [1,3,11] techniques are used to optimise the correspondence across N views. The initial shape estimate provides a starting point for the local shape optimisation. This approach enables improved reconstruction of local surface shape and correspondence among images from wide-angle views. Figure 2d illustrates the animation of the reconstructed virtual dancer model in a Venice street scene.

The current status of this technology for studio-based reconstruction of actor shape and movement is that the quality does not yet meet the requirements for photo-realism as set out above. Advances made in model-based shape reconstruction and movement capture indicate that computer vision will deliver video-quality 3D studio production, enabling virtual camera views and interaction within a virtual set. Recent advances in motion capture in this and other work [2] indicate that reliable and efficient capture of human movement can be achieved. Both shape and motion capture use prior generic models of human shape and anatomical structure specific to the particular application, which enables reliable reconstruction.

4. Conclusions

The development of vision-based techniques for human modelling, motion capture and analysis must start from the user requirements of the particular application. Automatic initialisation and analysis together with robustness are important problems for many potential mass-user applications. Many applications require trade-offs between automation vs. interactiveness and accuracy vs. realism which are central to the development of an adequate solution. Entertainment applications in games, virtual reality and film primarily require photo-realism as opposed to geometric accuracy. In addition, issues of cost and space are important in designing appropriate hardware capture systems. The model-based computer vision techniques developed in our research have enabled robust reconstruction from single-view and multi-view images for real applications. The use of strong application-specific prior models to constrain reconstruction helps in overcoming problems due to inherent visual ambiguities. Computer vision technology for human modelling which addresses application-specific requirements has a bright future of exploitation to address real mass-user applications.

Acknowledgements. This work was supported by UK EPSRC funding agency on Broadcast LINK project PROMETHEUS GR/M88075 and EPSRC Advanced Fellowship AF/95/2531. The author would like to thank the BBC, dancer Deborah Bull (deborahbull.com) and members of the author's research team Jon Starck and Joel Mitchelson for substantial contributions to the studio capture system.

References

1. Debevec P, Taylor C, Malik J (1996) Modeling and rendering architecture from photographs. In: Proceedings of the ACM SIGGRAPH computer graphics annual conference series, pp 11–21
2. Deutscher J, Davidson A, Reid I (2001) Automatic partitioning of high-dimensional search spaces associated with articulated body motion capture. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Hilton Head Island, SC, June 2001, pp 669-676
3. Fua P (1997) From multiple stereo views to multiple 3d surfaces. Int J Comput Vis 24(1):19–35
4. Fua, P. (2000). Regularised bundle-adjustment to model heads from image sequences without calibration data. Int J Comput Vis 38(2):153–171
5. Hilton A (1999) Towards model-based capture of a persons shape, appearance and motion. In: Proceedings of the IEEE international workshop on modelling people, Greece, September 1999, pp 37–44
6. Hilton A, Beresford D, Gentils T, Smith R, Sun W (May 1999) Virtual people: capturing human models to populate virtual worlds. In: Proceedings of the IEEE international conference on computer animation, Geneva, May 1999, pp 174–185
7. Hilton A, Beresford D, Gentils T, Smith R, Sun W, Illingworth J (2000) Whole-body modelling of people from multi-view images to populate virtual worlds. Visual Comput Int J Comput Graphics 16(7):411–436
8. Kanade T (1996) Virtualized reality: putting reality into virtual reality. In: Proceedings of the 2nd international workshop on object representation for computer vision ECCV, Cambridge, March 1996, pp 15–23
9. McLauchlan P, Jaenicke A (2000) Image mosaicing using sequential bundle adjustmnet. In: Proceedings of the British machine vision conference, Southampton, September 2000, pp 751–759
10. Moezzi S, Katkere A, Kuramura D, Jain R (1996) Reality modeling and vizualization from multiple video sequences. IEEE Comput Graph Applicat 16(11):58–63
11. Shan Y, Liu Z, Zhang Z (2001) Model-based bundle adjustment with application to face modeling. In: Proceedings of the international conference on computer vision, Vancouver, July 2001, pp 644–651
12. Starck J, Collins G, Smith R, Hilton A, Illingworth J (2003) Animated statues. J Mach Vis Applicat (in press)
13. Starck J, Hilton A (2002) Reconstruction of animated models from images using constrained deformable surfaces. In: Proceedings of the 10th international conference on discrete geometry for computer imagery, Bordeaux, France, April 2002. Lecture notes in computer science, vol 2301. Springer, Berlin Heidelberg New York, pp 382–391
14. Zisserman A, Fitzgibbon A, Cross G (1999) VHS to VRML: 3D graphical models from video sequences. In: Proceedings of the IEEE international conference on multimedia and systems, Florence, June 1999, pp 51–57