The use of a cast to generate person-biased video presentations

Dave Grosvenor, Susanne Klein
Media Technologies Laboratory
HP Laboratories Bristol

HPL-2007-11
January 30, 2007*

autorostrum, photo-video, rostrum slideshow, cast, person recognition, person identification, emphasis, emphasis statistics

This technical report was originally externally published as a defensive publication from an invention disclosure. However it was decided to make the paper easily available to the HP technical community.

A video presentation is generated using the specification of a "cast" to direct the emphasis of particular people.

It is a simple means of exploiting some additional semantic information obtained using object identification, and can easily be tuned to present many generic stories involving people without any deep semantic knowledge of the actual story.

The establishment of a cast allows people-oriented variations of video presentations that are understandable and controllable by a user. Manually establishing the "cast" is a powerful mechanism for controlling the presentation and provides a psychologically important step for a user to establish ownership over the generated presentation.

* Internal Accession Date Only

Approved for External Publication

The use of a cast to generate person-biased video presentations

Dave Grosvenor
Susanne Klein

26th July 2006

Introduction

This invention generates a video presentations using the specification of a "cast" to direct the emphasis of particular people or objects through the use of object recognition or identification techniques ([1][2][3][4][5]). The video presentations are composed from an input set of photographs and videos.

We will use the terms "person" or "people" and "face" to refer to a member or members of the cast and the primary view of an object, but it should be understood that the invention applies to the identification of arbitrary objects that can occur in the cast. This is done because people are the most important objects occurring in peoples photographs.

The invention generates aesthetically pleasing results for a non-expert user without a deep knowledge of the underlying story (such as is done by [7]). This contrasts with film and television production, where the shots and cuts are made by professional film-makers with an intimate knowledge of the underlying story. The invention provides a simple means of exploiting some additional semantic information obtained using object identification, and can easily be tuned to present many generic stories involving people without any deep semantic knowledge of the actual story.

This invention provides variations of video presentations that are understandable and controllable by a user. Manually establishing the "cast" is a powerful mechanism for controlling the presentation and provides a psychologically important step for a user to establish ownership over the generated presentation.

Emphasis in video and rostrum camera work

Video has insufficient resolution for the appreciation of the fine detail in a photograph whilst showing the whole photograph. But the fine details can be shown by panning and zooming a virtual rostrum camera about the photograph [6]. This generates a video from a still image. It is also a powerful creative technique for focusing a viewer's attention on part of the photograph or emphasizing a particular person or object. Such rostrum camerawork gives a lot of scope for creativity, because the duration of the video generated, the field of view of the rostrum camera, and the path it follows can be varied to produce many different effects.

Emphasis of a person in a rostrum video can be achieved by:

- Including the person in the field of view of the camera.
- Reducing the presence of competing or distracting objects
- Increasing the level of camera zoom.
- Increasing the time spent upon a particular face..
- The position of the face within the frame.
- The rostrum camera movement itself.
    - When the rostrum camera zooms into a region, a viewer's eye is drawn to the focus of expansion.

    - When the rostrum camera zooms out to reveal a larger field of view, attention starts within the initial field of view but will be drawn to the newly revealed objects.
    - Panning will introduce a bias for the point of attention to follow the direction of movement towards the new objects in the scene.

Creative manipulation of the emphasis produced by a video clip is more difficult.

- The field of view, duration, and camera movement in a video clip is relatively fixed.
    - The sound track makes it difficult to use cutting or fast-motion (or slow-motion) to change the duration.
    - There is generally insufficient resolution to permit substantial changes in the field of view or camera movement
- Object movement can now occur independently of camera motion. Object movement is a very powerful means of attracting attention, and an audience's attention is drawn to a moving object. This will de-emphasize any other non-moving objects within the field of view.
- But video can record complex and subtle interactions between people, and the audience's understanding is very sophisticated.

The cast

The "cast" might be determined automatically from analysis of the initial, or more powerfully it could be determined with some user interaction.

The specification of the "cast" directs the video editing algorithm to emphasize particular people (or objects) in the final presentation. The notion of a cast is used to indirectly control the appearance of the video-presentation. The cast is used :

- To identify the people to be emphasized.
- To identify relations between people by emphasizing particular groupings of people.
- To emphasize particular spatial configurations of people
- To specify the relative distributions of particular people and groups of people.
- To specify how the people and relations emphasized can vary throughout the video presentation. At the beginning of an event we might choose to equally weight each actor to introduce them, before choosing to emphasize the "star" actors. Similarly at the end of an event the entire cast might be shown again. These variations are usually the result of some stylistic parameter for the whole presentation.

These semantic observations of the cast of the video presentation will indirectly control the mix of shots (close-up, medium, long) used for particular people by measuring the emphasis given to particular people in the final album.

A simple measure of the emphasis placed upon a particular person is the sum of the area of their face's whenever they occur in the final video presentation. More complex measures would

- Weight the sharpness and quality of the face image.
- Spatially weight the face according to its position within the final video frame.
- Take into account the relative size of other faces on the page.

Analysis

Techniques of object recognition and visual similarity ([1][2][3][4][5]) are used to analyze the input photo-set to measure the emphasis on different people and groupings of people. This analysis gathers statistics of the people or objects identified in the original photo-set recording:

- The combinations of people that occur together.
- The emphasis placed upon particular people by their size and position in the original photo
- The introduction of newly identified people into the presentation.

- The importance of pictures near the start and finish of scene or event boundaries within the original photo-set.
- The unusual pictures in the photo-set.
- The sets of visually similar pictures.

This analysis can use metadata from the larger photo-collection containing it. This makes it easier to both recognise people already known in the collection, and identify useful relationships between these people (such as husband, wife, parent, etc…) that could affect the cast.

This analysis provides a context for the photo-set. The cast assigns roles for people occurring in this context for the photo-set. But the statistics for the object emphasis produced by the cast need not reflect the original context. The emphasis statistics provided by the original photo-context can be modified by either stylistic or manual controls. In the extreme, the cast can produce emphasis on particular actors that is independent of the emphasis present in the input photo-set.

Emphasis driven video editing

Once the cast has been determined we have specified the desired emphasis. The video editing algorithm has to generate a video presentation with an acceptable fit to the desired emphasis. i.e. as directed. This requires some form of optimisation-like search through the space of potential edits and the use of the emphasis measure to compare the desired emphasis (given by the cast) with that created by a particular editing composition.

The flexibility provided by the many different ways of generating a rostrum video clip from each photograph creates a combinatorial explosion because any of these could be joined to any of the variants of the following clips. Thus any practical implementation of the video-editing algorithm needs to use various assumptions to reduce the search space.

In particular, we assume that a video-style has been determined that can be used to restrict the space of creative designs explored. The video-style is responsible for setting the look and feel of the whole presentation. It controls the:

- Pace and rhythm of the video presentation
- Rate of cutting and transitions (dissolves, fades)
- Synchronisation with the music.
- Use of both stock footage and graphical animation

Most importantly for our purposes, this video-style must characterise the potential sequences of camera motions that can be combined together which are consistent with the video-style.

We do this using the notion of a video-template. It combines a group of rostrum camera motions (made from a number of photos) and video clips into a small video clip. A whole video presentation will be built using a sequence of such video templates.

The video-style characterises a set of templates that can be used together to construct the video presentation. The set of video-templates would have been designed by an expert to be consistent with each other and yet different enough to introduce variation.

Video-template

The video-template performs particular rostrum camera movements on the input photos, and combines these into a small video clip. A video-template:

- Characterises the rostrum camera movement that is performed on each input photo.
- The actual rostrum camera movements are only crudely characterised by the regions of interest and direction of movement. This freedom allows different photos to be used in the

template. The final rostrum camera path would be modified by image analysis to ensure correct cropping and sensible end-points for panning and zooming motions.

- Characterises the transitions between the different video clips combined (including the rostrum videos)
- Enables the calculation of the emphasis given to the cast.

The video-template encodes many creative design decisions:

- The combination of camera motions in different directions. .
- The combination of different camera speeds.
- The variation in the duration of the rostrum video clips.
- The use of contrasting camera motions to make a viewer aware of a cut, and at other times the cut will be disguised by using similar motion.
- Arranging adjacent camera movements to have common points of attention to transfer attention from one object to another.

References

1. “Pattern Classification”(2nd ed.) by Richard O. Duda, Peter E. Hart and David G. Stork, Wiley Interscience, 680 pages ISBN: 0-471-05669-3
2. “Dynamic Vision: From Images to Face Recognition”
By Shaogang Gong, Stephen McKenna, Alexandra Psarrou , 364 pages, Imperial College Press, 2000.
3. “*State-o-fthe-Art in Content-Based Image and Video Retrieval”,* Veltkamp, R. C., Burkhardt, H., and Kriegel, H.-P., editors (2001. Kluwer Academic publishers.
4. “*Image Databases: Search and Retrieval of Digital Imagery*”, Castelli, V. and Bergman, D., editors (2002). John Wiley & Sons, Inc.
5. “Special issue on content-based multimedia indexing and retrieval”, Djeraba, C. et al. (2002).. *IEEE Multimedia Magazine*, 9(2):18.60.
6. “Auto-rostrum patent – a method for the display of a digital image” , ~D.A. Grosvenor, S.P. Cheatle, US20020118287
7. “Muvee autoproducer”, Automatic video presentation generation from music, photographs and video http://www.muvee.com/ .