AN EMBEDDED SUGGESTIVE INTERFACE FOR MAKING HOME VIDEOS

Brett Adams, Svetha Venkatesh

Department of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{adamsb, svetha}@cs.curtin.edu.au

ABSTRACT

This paper describes a novel suggestive interface embedded in a smart camera prototype aimed at aiding home movie makers. We focus on the problem of generating shot capture suggestions suitable to the user's filming context, intended audience and style, and formulate a novel aesthetic measure by which to judge proposed suggestions. Tight coupling between media and software allows the aesthetic measure to be sensitive to previous footage captures, including those taken without the system's prompting, in a manner allowing flexible, end-to-end migration of the authoring task from user to machine. An approximate method is used to find timely, near-optimal solutions to the aesthetic measure. Qualitative evaluation in the form of a user study shows it to be a promising approach to the flexible home movie authoring context.

1. INTRODUCTION

Wide scale embedding of computation in handheld devices capable of media capture (e.g. images, video) has created the potential for providing authoring support previously limited to an offline, after-the-event desktop phase. In fact, the coupling of digital media and computation for the duration of the media creation lifecycle, from capture onwards, creates entirely new possibilities: at the very least, media capture instances can be supported by timely feedback built upon per user profiles, context awareness (e.g. biological, locational, social, historical), and estimates of complex aesthetic or narrative features (e.g. visual composition, global fitness to style, rhythmic impact). Together these form an analog to existing camera feedback about simple attributes like focus or lighting, but extended in N infinitely redefinable dimensions by the power of the digital domain.

In other words, mobile media capture devices provide the ability to migrate tasks performed manually to automatic computation, in a flexible manner. For amateur home media creation, particularly video, flexibility is key, as there are as many user preferences, goals, and ways of working as their are users. Each will wish to migrate different subsets of the media creation task to computation. E.g., some want tighter control over fine, frame-level edits, while others apportion more energy to gathering as much footage of an event as possible.

Suggestive interfaces [1] are an effective means for providing flexible task migration. Typically software agents, each with knowledge of a particular task within the domain, suggest possible completions to complex activities, based on observation of user activity. E.g., [2] use a suggestive interface for a 3D drawing application. The user hints at the desired operation by highlighting related geometric primitives in a scene, the system infers possible operations and communicates these as thumbnails (e.g. cut the corner of a polygon), and the user confirms their intention by choosing from the offered suggestions. For amateur media creation this work–suggest–confirm protocol has an additional, valuable outcome: it inherently creates semantically rich annotation, thus supporting a chief concern of the digital revolution, media search.

Related work on software-assisted video creation is either limited to post-production (e.g. [3] and a host of commercial solutions), or is not coupled to the raw medium itself during capture (e.g. [4]). Consequently these approaches are unable to give real-time feedback about the quality of captured media, when reattempt is possible.

We present eMediaTE (embedded Media To Everyone), a completely mobile software-assisted home video authoring solution, deployed on a prototype smart camera. Unlike previous work [5], eMediaTE generates shot suggestions in real time by finding near-optimal solutions in a novel composite aesthetic measure space, based crucially on the user's intention and *all previously captured media*. eMediaTE thus supports unplanned video capture, providing a good fit to amateur video creation practice, and yields the following significance:

- Improved video quality - via supervised capture,
- Bounded authoring - the video capture act *is* the authoring activity, leaving 'no more to do' akin to photo capture,
- Semantically rich annotation - an important feature as video capture is increasingly viewed as being integrated with other media types (e.g. photos, blogs), and in a workflow that includes sharing, archival, reuse [6], and
- A platform for integrating context awareness - a fast growing focus of research in mobile computing.

Project funded by the Australian Research Council, and Curtin University of Technology Fellowships Scheme.

The remainder of the paper will provide an overview of the implemented system, formulation of a history-dependent aesthetic measure, and qualitative experience with the system via user feedback.

2. SYSTEM OVERVIEW

eMediaTE has been implemented in C++ and deployed on a prototype smart camera consisting of a Sony Vaio U71 running Windows XP harnessed to a DV Camera. The system is wielded in one hand, with the Vaio forming the 'smart' viewfinder of the DV Camera. Interaction with eMediaTE is via the Sony's touchscreen, and optionally ViaVoice. Figure 1a. is a photo of the system. A typical video creation session with the system involves a small number of simple interactions:

Elicitation of the user's filming context: Up to three questions are posed to the user, about where they are, and who and what they are interested in filming. Nouns are parsed from the voice or text input and become labels for entities at the filming site. Entities with labels that are proper or singular nouns are assumed to be unique, and the remainder are assumed to be non-unique. E.g., *George* and *a dog* would both be assumed unique, whereas *people* or *trees* would not. The user is optionally able to indicate the desired nature of the final video, or its communicative intent. Currently, a choice of several genres–being readily understood by the amateur–are available, such as Action, Drama, Experimental French Cinema, Sitcom, and Cooking Show. Genres are implemented as a mapping to a more fundamental representation of video style and communicative intent, and are thus open to a vast array of configurations. The initiation of a fresh user elicitation is by explicit request from the user, but scope to initiate it passively is a simple extension, via detection of changes to location (e.g. via GPS-based location services) and/or time.

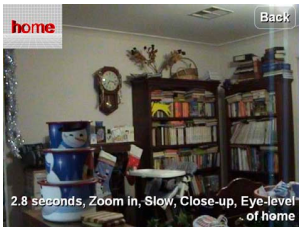
Issue suggestions: Following elicitation, a shot capture, or at the explicit request of the user, a small number of suggestions for shots to be captured are displayed. A suggestion consists of a fixed configuration of a number of primitive cinematic elements, such as subjects, camera motion, duration, framing type, and so on. Suggestions are chosen that maximize an aesthetic measure dependent on the target video genre, previously captured footage (including footage not taken in response to a suggestion), and proposed suggestion configuration. Offered suggestions at a given time are constrained to be sufficiently different from each other, yet near-maximal, to increase the chance of at least one being appropriate.

Capture shot: Having selected a shot suggestion to attempt, the user records footage. Upon completion of the shot recording, the user is asked to indicate whether or not they thought they succeeded. Their response, either yes or no, translates directly to the attached inferences about metadata of the footage. For a 'yes' response, crucially, annotation of a nature that is difficult to arrive at automatically, is obtained. E.g., regarding framing type (close-up, long shot) and subject presence (George is present in this shot). No user-derived metadata

(a) eMediaTE on prototype smart camera.

(b) Elicitation

(c) Issue Suggestions



(d) Capture Shot

(e) Preview

Fig. 1. eMediaTE

inferences are made for a 'no' response. At any stage the user can simply record, as with a traditional camera, with the resulting footage being marked as impromptu. Impromptu footage is processed to obtain automatic metadata (e.g. camera motion detection, and estimates of framing type).

Preview: At any point during filming, the user may view raw shot captures, or an automatically edited preview of the video thus far. Automated editing includes the addition of appropriate shot transition effects and audio overlay, as well as fine-grained decisions, such as choice of frame ranges that contain the desired motion type for a given shot.

Archive/share: Video compositions, together with metadata about footage content and editing decisions, can be exported to a home archive or uploaded to a video blog for immediate sharing via WIFI. Export is predictably performed upon completion, but conceivably also *during* an event. In that case partial videos can be used to supply closer to real-time coverage of a live event to the 'distant interested' (e.g. a wedding).

3. FORMULATION OF CAPTURE HISTORY-DEPENDENT AESTHETIC MEASURE

The abovementioned aesthetic measure can be thought of as analogous to Birkhoff's aesthetic measure for polygons [7]. In this case we seek a measure of desirability of a proposed shot suggestion, in light of the desired genre for the video, narrative point, and previously captured footage. E.g., for a video of desired genre 'Action', a very simple aesthetic measure might favour shot suggestions with high motion levels or short duration, rather than those that are static or of longer duration. More formally, our aesthetic measure is defined as:

$$AM(s, g, \tilde{a}) = d \in [0:1] \quad (1)$$

Where s is a shot suggestion, g is the desired genre for the video, and $\tilde{a}$ is all previous shot captures. The function range is from 0, least desirable suggestion, to 1, most desirable.

AM can be further broken into a weighted sum of specialized sub-functions,

$$AM(\theta) = \omega_1 AM_1(\theta) + \omega_2 AM_2(\theta), \cdots + \omega_N AM_N(\theta) \quad (2)$$

Where each AM_i is defined similarly, and $\sum_{i=1}^{N} \omega_i = 1$.

A number of video aesthetics have been modelled and implemented as evaluator sub-functions, AM_i, in eMediaTE. Some of these film techniques have been used in previous work, however their formulations as aesthetic measures are new, to fit with the current 'fully impromptu' and 'media-coupled' setting. The reader is referred to [5] for detailed descriptions of how they are manipulated in film, their aimed-for effects on the viewer, together with mappings between particular genres and techniques employed.

Due to space constraints, we will discuss only the evaluator specializing in the filmic technique of visual approach, AM_{VA}. In film, visual approach to a scene aims to initialize the audience's 'mental map' of a scene's constituents, their spatial layout and inter-relationships. It tends toward being either Deductive, moving from wider shot types to tighter, or Inductive, consisting mostly of close framing types with perhaps a resolving wider shot following. These types effectively manipulate the amount of scene *context* the user can grasp, which in turn effects whether the film feels more or less intense. Different genres and/or styles thus use it to evoke a certain response from the viewer.

eMediaTE models two aspects of visual approach in order to implement an evaluator AM_{VA}. The first is an estimate of an average viewer's apprehension of the current scene's parts and their interrelationships, termed *context*:

$$cxt(\tilde{f}, \tilde{p}) = \sum_{j=1}^{n} A(f_j) e^{\left(\frac{j-n}{\tau(p_j)}\right)} \in [0:1] \quad (3)$$

Where the parameters are $\tilde{f}$, a hypothetical or actual sequence of framing types (e.g. Long shot, medium shot etc.) of previous shots, and $\tilde{p}$, the corresponding sequence of filmed subjects as a percentage of the total subjects in the scene. The weighted sum of shifted exponentials on the right hand side can be viewed as successive 'drops' of context fed to the viewer at each shot up to the current n, parameterized by a beginning amplitude $A(f_j)$ and a time constant $\tau(p_j)$. $A(f_j)$ increases with wider framing types (e.g. Extreme long shot being the widest), and $\tau(p_j)$ increases with greater percentage of subjects covered in a shot. $cxt()$ is clipped to a ceiling of 1. The second aspect of visual approach modelled is subject *coverage*, which is simply a histogram of subjects covered up to the current shot, $cvg()$. Finally, the evaluation AM_{VA} can be calculated using the sum of two distance measures,

$$AM_{VA}(s, g, \tilde{a}) = 1 - (D_{cxt}(X_{\tilde{a}+s}, X_t) + D_{cvg}(V_{\tilde{a}+s}, V_t)) \quad (4)$$

Where X_{a+s} is the histogram of $cxt()$, whose bins are shots up to and including the proposed suggestion, and similarly X_t is the histogram of target $cxt()$ values. $D_{cxt}()$ is the ordinal histogram distance measure of [8], normalized by the best and worst possible distance for the given g and $\tilde{a}$. Similarly, $V_{\tilde{a}+s}$ is the histogram of subject count up to and including the proposed suggestion, and V_t is a set of equally desirable target histograms of subject count up to the current shot. The two distance measures are subtracted from 1 to yield a desirability measure ranging between 0 and 1.

Each AM_i follows a similar pattern of modelling aspects of interest, calculating distance between the proposed suggestion and the best possible, and transforming this into an evaluation of the suggestion's desirability. Other implemented AM_is include: Scene orchestration AM_{SO}, responsible for evaluating a suggestion's ability to further cover the scene's subjects, while manipulating $cxt()$ and observing correct emphasis on people or objects, and external 'plot' or internal 'character'; Visual tempo AM_{VT}, for manipulating the amount of information thrust at the viewer; and Film sense AM_{FS}, for devaluing proposed suggestions that are either combinations of cinematic operations too difficult for an amateur, or are otherwise cinematically nonsensical (e.g. a static close-up of many people). The aesthetic measure can be extended in its expressiveness by simply adding new sub-functions.

4. FINDING NEAR-OPTIMAL SOLUTIONS TO AESTHETIC MEASURE

Ideally, shot suggestions issued to the user by eMediaTE at any stage will be maxima in the aesthetic measure AM. Expressed formally, we desire any suggestion s_0:

$$s_0 = \max_s AM(s, g, \tilde{a}) \quad (5)$$

We note that the search space of all possible suggestions is intractably large.[1] Consequently an approximate method is required to find near-maximal solutions to Equation 5. The heterogenous nature of aesthetic measure sub-functions renders the parent function AM discontinuous, and lacking any clear notion of direction. Simulated annealing has been found to perform well in such conditions and has been chosen here.

Additionally, eMediaTE must generate a *number* of suggestions that are all near-maximal, but sufficiently different from each other so as to maximize the possibility that at least

[1] A suggestion consists of an ordered set of subjects, together with a number of elements that take on a large, albeit discrete, range.

one of them will be deemed appropriate by the user. Thus the suggestion generator feeding the annealing process is also able to be constrained to generate suggestions at least a certain Manhattan distance from a set of reference suggestions. I.e., the first s_0 is found, then another a minimum distance from it, s_1, and another distanced from the set $\{s_0, s_1\}$.

Figure 2 is an example of the convergence of AM for a single suggestion generation. The annealing schedule used is geometric with a 100 temperature decrements with ratio of 0.97, and 1000 samples at each temperature. Initial temperature is 0.5. Time to anneal is sub one second. It can be noted there are many suggestions near the theoretic maximum of Equation 1. The occurrence of values of AM so near this maximum will decrease with the addition of AM_i's, particularly those with competing criteria for desirable suggestions.

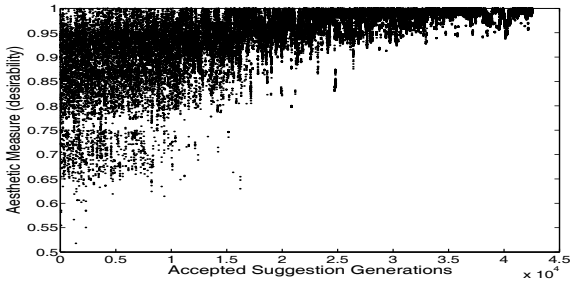


Fig. 2. Example convergence of suggestion desirability.

5. EVALUATION

An informal user study was undertaken to gain a qualitative estimate of the usability of the suggestive interface and smart camera prototype. 7 users were given a brief demo of making a video with the system, and asked to make a short movie with it. All bar one had not used a DV camera before. An interview was conducted with each user upon completion of their task, and in particular focused on their attitude toward the dynamically generated shot suggestions. In response to the question "Would you use this camera again?" all users responded positively. Importantly, all demonstrated the ability to navigate the selection, capture and verification of shots.

The issue that uniformly generated the most discussion was filming spontaneous action. 3 of the 7 users said they would feel more comfortable filming events or objects that were not highly dynamic, such as parties with many subjects who are mobile. This remains one of the biggest challenges to successfully using suggestive interfaces for impromptu filming. Two critical components were noted: speedy elicitation of filming context and integrating captures taken in response to suggestions versus impromptu captures. eMediaTE's option for voice input is aimed at alleviating this time-squeeze, but unfortunately the user study was unable to include a voice training phase, and hence the onscreen keyboard was used instead. The second need might be largely addressed by reversing the default mode of operation. Currently, eMediaTE's interface assumes the user will be predominantly using suggestions rather than impromptu filming. Alternatively, the interface could initially appear to be a 'dumb' camera, which consequently passively flags requests for context information and the subsequent availability of suggestions. In that case, for a user who has time to supply information and peruse suggestions, the capture will follow the currently implemented path. But for users who are frantically trying to 'record' everything, it will fail soft. I.e., the interface should assume the user is time-starved, and let them engage with the suggestive interface at their own pace.

While some users were "happy to take suggestions," others anticipated the desire for communication of the *reason* for the particular suggestions offered. A natural solution to this using the aesthetic measure would be for the sub-functions to include textual interpretations of a given suggestion's role in their evaluation of it, and then publish this to the user.

6. CONCLUSION AND FUTURE WORK

We have presented a novel suggestive interface embedded in a smart camera prototype aimed at aiding home movie makers. Shot suggestions are generated by means of a novel aesthetic measure targetted at video composition, together with an implementation for obtaining timely near-optimal solutions. An informal user study demonstrated both the system hardware and authoring concepts were understood by beginners, with the chief challenge noted as being improved accommodation to time-starved users in dynamic environments. Future work will include: improving the parameters for $cxt()$ via learning, as training data is plentiful, in the form of shooting scripts (containing directions for framing type and subject) together with genre groundtruth; more evaluators, such as rhythm; further AM-sourced help overlayed on the viewfinder in real time, such as desirable visual compositions, an opportunity provided by the smart camera platform. Integration of the media capture capabilities with a presentation environment for heterogenous media types (video, images, audio, text), on the device itself, augmented by context awareness (e.g. location history), presents an exciting mobile media paradigm.

7. REFERENCES

[1] N. Diakopoulos and I. Essa, "Supporting personal media authoring," in *ACM Int. Conf on Multimedia*, Nov. 2005.

[2] T. Igarashi and J.F. Hughes, "A suggestive interface for 3d drawing," in *ACM symposium on User interface software and technology, Orlando, Florida*, November 2001.

[3] X.-S. Hua and S. Li, "Personal media sharing and authoring on the web," in *ACM Int. Conf on Multimedia*, Nov. 2005.

[4] B. Barry, *Mindful Documentary*, MIT, Ph.D. thesis, 2005.

[5] B. Adams and S. Venkatesh, "Situated event bootstrapping and capture guidance for automated home movie authoring," in *ACM Int. Conf on Multimedia*, Nov. 2005.

[6] F. Nack, "Capture and transfer of metadata during video production," in *ACM Workshop on Multimedia for Human Communication - From Capture to Convey*, Nov. 2005.

[7] G.D. Birkhoff, *Aesthetic Measure*, Harvard Uni. Press, 1933.

[8] S.-H. Cha and S.N. Srihari, "On measuring the distance between histograms.," *Pattern Recognition*, vol. 35, no. 6, pp. 1355–1370, June 2002.