INJECTION, DETECTION AND REPAIR OF AESTHETICS IN HOME MOVIES

Brett Adams, Svetha Venkatesh

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

ABSTRACT

This paper details the design of an algorithm for automatically manipulating the important aesthetic element of video, visual tempo. Automatic injection, detection and repair of such aesthetic elements, it is argued, is vital to the next generation of amateur multimedia authoring tools. We evaluate the performance of the algorithm on a battery of synthetic data and demonstrate its ability to return the visual tempo of the final media a considerable degree closer to the target signal. The novelty of this work lies chiefly in the systematic manipulation of this high level aesthetic element of video.

1. INTRODUCTION

Amatuer videographers often lack the time or expertise required to fashion video compositions that faithfully communicate their experience of an event. Related work [1] and, increasingly, commercial offerings [2], seek to lower the barrier for home movie makers by automating the tedious and time consuming editing phase, crucially without requiring manual annotation. They do so by leveraging general metrics about the 'interest' of footage calculated from objects and camera motion, and the presence of specific objects such as faces etc., together with editing rules. The problem with this approach is twofold: The resulting edits are not content-sensitive, as no semantically rich metadata exists to inform the process, and secondly the content and cinematic parameterization of the footage the computer receives may already severely constrain its aesthetic or communicative potential (e.g. style). [3, 4] and others have noted that computer-assisted solutions to this problem require embedding technology at/before capture. We agree this timely presence is vital, not only to inform users regarding choice of subject and cinematic parameters from the myriad options available, but *also* to enable picking up of the pieces, so to speak, that result from the contingent nature of the capture context the amateur videographer often works within. We have implemented a prototype system aimed at satisfying these needs. Refer to [5] for a comprehensive discussion of the system.

This paper is restricted to new work on the second of the above requirements: automatic content-sensitive injection and repair of aesthetic elements of video, in particular the element of *visual tempo*. Tempo is dependent on the shot length and motion characteristics of footage. We demonstrate that it can be manipulated toward a target in a manner sensitive to other properties of video. E.g. tempo at a point in a video may be raised or lowered through the search for nearby footage that meets particular motion characteristics. Additionally, in order to test the algorithms presented here, we have developed methods for generating synthetic data subject to parameters in a repeatable manner, another novel aspect of this work. These algorithms have been developed within our existing media creation framework, but it is stressed that they are not limited to this context, but are applicable wherever the user or system can specify a target tempo signal at which to aim. Such technology is vital to enabling aesthetics-aware, (semi-)automatic media creation workflows, which to the authors' knowledge has not been attempted in a systematic fashion.

2. INTEGRATED MEDIA CREATION

Firstly, what are media aesthetics and why are they significant? [6] defines applied media aesthetics in part as: *The manipulation of media elements, such as light and colour, two and three-dimensional space, time and motion, and sound ... for synthesis that "clarifies, intensifies, and effectively communicates an experience."*

[7] outlined the goals and methods of the nascent area of Computational Media Aesthetics (CMA), the algorithmic application of these ideas to existing media, and subsequent work [8] has demonstrated the automatic detection of aesthetic elements. But can these same aesthetic elements can be practically *injected* into amateur video in the case of video synthesis?

Injecting and repairing aesthetics requires an organic approach to media creation, and hence we have created an Integrated Media Creation Environment (IMCE) that supports the creation of video through the entire workflow, from conception to finished product, and on through repurposing.

0-7803-9332-5/05/$20.00 ©2005 IEEE

The reader is referred to [5] for a formal treatment of IMCE, not possible here due to space restrictions. A qualitative description of IMCE and example must suffice to provide the necessary background to the core discussion of this paper.

There are a number of phases to creating a video with IMCE. Initially some conception of the occasion to film is obtained, including structural information such as sub-events and parts (both human and non-human), and optionally narrative attributes (e.g. this event is the climax). E.g. a birthday party may include an event cutting the cake, participants wife and friends, and objects such as the cake and presents. This information may be created by the user, obtained from a library, or some combination thereof, potentially augmented by semantic net techniques [9]. Methods of bootstrapping the process by collecting this information on-the-fly in the field are under investigation. Next IMCE generates a shooting list of shots with desired cinematic parameters to be filmed. These *shot directives* consist of P parameters, including subject, motion level and direction, framing type, and so on, thresholded from easier to harder according to user expertise level. Shot directives are delivered to the user in the field via a palm device, through which he also implicitly creates a record of the capture history as he interacts with it to indicate success or failure at filming given shot directives and impromptu shots. Thus instead of engaging in common technical and aesthetic pitfalls, e.g. panning all over the scene; using the zoom excessively; or placing every subject at the center of the frame [10], the shot directives guide the user into a harvest of footage pregnant with aesthetic potential and semantic annotation, such as angles to communicate the complexity of the party; closeups to reveal emotion; long shots to contextualise events; and motion and shot timings to heighten excitement. Following filming, the process moves to the desktop [1], where the capture record enables mating of shot directives with corresponding raw footage, metadata without manual annotation. IMCE then automatically edits the raw footage into a movie, and it is during this process that discrepancies between the desired aesthetic elements and those achieved are detected, and repair is attempted. E.g. a missed shot directive may cause the party climax to not be adequately emphasized by a rise in tempo, as originally conceived by IMCE, and thus it takes corrective action to create the rise using material that *has* been captured. The remainder of this paper focuses on this process for the aesthetic element of visual tempo.

3. PROBLEM FORMULATION

Tempo is defined as "perceived speed of a [filmic] event" [6], hence it is also termed "subjective time." The visual component of tempo is computed via the algorithm of [8]:

$$T(n) = W(s(n)) + \frac{m(n) - \mu_m}{\sigma_m} \tag{1}$$

Where $s(\cdot)$ refers to shot length in frames, $m(\cdot)$ to average motion magnitude of a shot, and n to shot number. $W(\cdot)$ is a weighting function that models the impact of (non-normally distributed) shot length on visual tempo. μ_m and σ_m are the motion mean and standard deviation respectively. Importantly, $T(\cdot)$ is also smoothed with a Gaussian of σ to reflect the *inertia* of visual tempo.

Let the sequence of shot directives generated by ICME be $\tilde{s}_i^T$, where i indicates the sub-event or *scene*, $i : 1 \rightarrow N$. These are shots IMCE has asked the user to attempt to film. Each shot directive s_{ij}^T has P properties including desired duration $s_{ij}^T.duration$ and motion magnitude $s_{ij}^T.motion$, which are pertinent to the calculation of visual tempo. Following filming, there exists another sequence of *realized* shot directives, shot directives aggregated with actual captured footage, $\tilde{r}_i^C$. Realized shot directives have two properties that shot directives do not: captured raw footage $r_i^C.raw\_footage$, and a frame range selected from this to be rendered into the final movie $r_i^C.selected\_footage$. A depiction of $\tilde{s}_i^T$ and $\tilde{r}_i^C$ can be found at the top of Figure 1a. An initial rough cut of the movie is generated by setting the selected footage property for all realized shot directives to a frame range that best matches the desired properties of the shot directive. Thus target visual tempo, what would be obtained *if* the target shots were captured correctly, can be calculated from $\tilde{s}_i^T$, and achieved visual tempo in this rough cut from $\tilde{r}_i^C$. Example plots of $T(\cdot)$ appear in the middle of Figure 1a. Deviation of $T(\tilde{s}_i^T)$ from $T(\tilde{r}_i^C)$ will cause IMCE to attempt to automatically repair the aesthetic element of visual tempo to recover the intended signal.

$$\tilde{r}_i^R = \begin{cases} \tilde{r}_i^C & : \text{if } T(s_{ij}^T) - T(r_{ij}^C) < \tau, \forall j : 1 \rightarrow |\tilde{r}_i^C| \\ R(\tilde{s}_i^T, \tilde{r}_i^C) & : \text{otherwise} \end{cases} \tag{2}$$

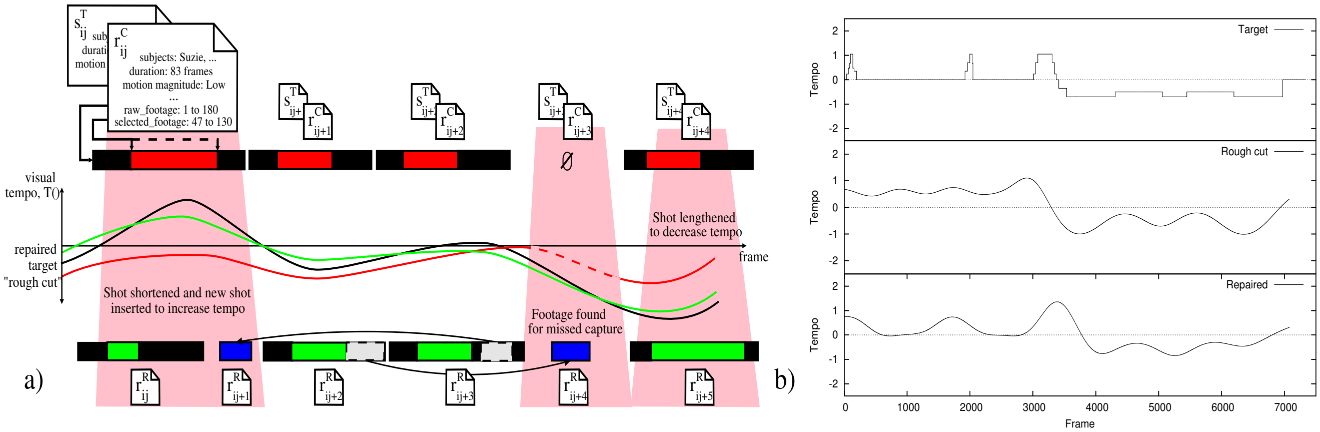
Where $\tilde{r}_i^R$ is the shot sequence to be rendered into the final movie $i : 1 \rightarrow N$, τ is a threshold of deviation of actual visual tempo from target, and $R(\cdot)$ is the visual tempo repair function, defined as:

$$R(\tilde{s}_i^T, \tilde{r}_i^C) = \min_{E(\tilde{r}_i^C)} T(\tilde{s}_i^T) - T(E(\tilde{r}_i^C)) + C(E(\tilde{r}_i^C)) \tag{3}$$

Where $E(\cdot)$ is an edit function which manipulates a shot sequence by means of insertions and deletions, and shifts and resizing of the frame range $r_{ij}^C.selected\_footage$, and $C(\cdot)$ is a cost function reflecting the potential detrimental impact of an edit to movie structure, such as continuity. Three deviations above τ, and action taken in response, can be noted in Figure 1a in the vertical bands, with the overall repaired shot sequence appearing below.

In other words, repair of the visual tempo of the manifest movie will be attempted when it strays beyond an acceptable threshold from the original target tempo.

[1] Due to implementation issues at present. In the near future, with rising mobile computing power and camera integration with IMCE, the video creation process could be completely decoupled from the desktop.



a)
s_{ij}^T
r_{ij}^C
subjects: Suzie, ...
duration: 83 frames
motion magnitude: Low
...
raw_footage: 1 to 180
selected_footage: 47 to 130
s_{ij+1}^T r_{ij+1}^C
s_{ij+2}^T r_{ij+2}^C
s_{ij+3}^T r_{ij+3}^C
s_{ij+4}^T r_{ij+4}^C
Ø
visual tempo, T()
repaired
target
"rough cut"
frame
Shot shortened and new shot inserted to increase tempo
Footage found for missed capture
Shot lengthened to decrease tempo
r_{ij}^R r_{ij+1}^R r_{ij+2}^R r_{ij+3}^R r_{ij+4}^R r_{ij+5}^R
b)
Target
Rough cut
Repaired
Tempo
2
1
0
-1
-2
0
1000
2000
3000
4000
5000
6000
7000
Frame

Figure 1: (a) Operation of $R(\cdot)$ on scene i: target shot directives $\tilde{s}_i^T$, rough cut $\tilde{r}_i^C$, and repaired shot sequence $\tilde{r}_i^R$, which is rendered; (b) $\tilde{s}^T$, $\tilde{r}^C$, and $\tilde{r}^R$, for a good synthetic capture, and no extra successes.

4. DESIGN OF $R(\cdot)$

In order to minimize $T(\tilde{s}_i^T) - T(E(\tilde{r}_i^C)) + C(E(\tilde{r}_i^C))$, all possible edits $E(\cdot)$ on the given shot sequence $\tilde{r}_i^C$ must be enumerated. Edits may alter both of the variables input to the tempo function, namely average shot motion and shot length. The motion of a given shot will change unpredictably with changes to the shot length or footage range chosen, so an analytical solution to $R(\cdot)$ is not possible. Moreover, the search space of all possible edits is vast. To enable a brute force search a number of domain heuristics are used to trim the search space to a manageable level.

4.1. Searching all edits, $E(\tilde{r}_i^C)$: constraints

⋆ Edits that insert footage from a different scene to that being repaired are excluded, as damage to movie continuity is more likely if footage from different locations or narrative time periods is mixed.[2]

⋆ No duplicate shots or jump cuts, consecutive shots drawn from the raw footage of a single shot, are allowed.

⋆ Edits to the shot sequence that do not create a true cut in the resulting movie are invalid. E.g. simply splitting a realized shot r into r_1 and r_2 without interposing footage between does not form a new shot in the movie.

⋆ When selecting new footage from within raw captured footage by shifting or resizing $selected\_footage$, a granularity of 0.5s is used. Precision greater than this leads to diminishing returns in terms of audience appreciation.

⋆ Only shot sequences that preserve *all* shot directive properties, not only those pertinent to tempo, (e.g. framing type or camera mounting type), are offered as candidates by $E(\cdot)$. Many aesthetic elements have been embedded in $\tilde{s}_i^T$; This policy preserves them whilst attempting to repair visual tempo.

⋆ Every shot r_{ij}^C may be replaced by up to two shots in $\tilde{r}_i^R$. Much can be done to alter tempo with one other shot, and in the case where a series of target shots deviate from the intended tempo, a shot may be inserted for each, with the overall effect being the insertion of multiple shots.

4.2. Searching all edits, $E(\tilde{r}_i^C)$: algorithm

For a given scene i different search paths are taken by $E(\cdot)$, leading to different candidate shot sequences returned, depending on the nature of the difference $T(s_{ij}^T) - T(r_{ij}^C)$.

If the tempo of shot r_{ij}^C is higher than the corresponding target shot s_{ij}^T, then $E(\cdot)$ will first try to lengthen the duration of r_{ij}^C from the currently set duration (remembering that tempo is inversely impacted by shot length), whilst observing any constraints detailed in the preceding section. This is possible where there exists successfully captured footage in excess of the originally desired duration, which is often the case. Figure 1a has an example of this at r_{ij+4}^C. If this fails to drop the tempo difference below τ, $E(\cdot)$ attempts to create a new shot from unused footage of nearby shots (e.g. excess footage, or where there are multiple successful attempts for a shot directive), with the aim being to find a two shot combination, the original r_{ij}^C and the new shot, that has a lower average tempo. If a winning combination is found, with the lowest difference, the new shot is inserted into $\tilde{r}_i^R$.

Conversely, if the tempo is lower, and hence needs to be raised, $E(\cdot)$ attempts to decrease the length of r_{ij}^C. Failing this, r_{ij}^C is split into two shots, necessitating the creation of a new shot. If the difference still exceeds τ, a new shot is created from a nearby shot, similar to the case above, in order to achieve a higher average tempo with the two shots. Figure 1a has an example of this at r_{ij}^C.

Finally, the only remaining situation consists of a target shot s_{ij}^T with no successfully captured footage at all, i.e. $r_{ij}^C.raw\_footage = \emptyset$. In this case, r_{ij}^C is replaced by a new shot whose footage is drawn from unused footage from a nearby shot. $E(\cdot)$ is here trying to match the target tempo of the original target shot s_{ij}^T with the new shot. Figure 1a has an example of this at r_{ij+3}^C.

[2] There are situations where this would be allowable, e.g. if parallel scene orchestration is being employed, but at present the component of IMCE responsible for scene orchestration does not include this technique.

5. EXPERIMENTS AND DISCUSSION

Evaluation of $R(\cdot)$ has two important facets: its ability to reduce the difference between the desired and achieved tempo signals, and the acceptableness of the edits employed to do so. The latter would require a large number of real filming projects to be undertaken, which is lacking at present. However, IMCE has been used to create a small number of videos, and [5] contains a user study structured as a Turing Test of the post-production phase. It should be noted many of causes of jarring edits are ruled out by virtue of the constraints applied in Section 4.1.

The second aspect of evaluation, closing of the tempo difference $T(\tilde{s}_i^T) - T(\tilde{r}_i^C)$, has been performed with the aid of synthetic data. Synthetic data is useful for the ability to specify groundtruth in a repeatable manner. Synthetic captures, i.e. $\tilde{r}^C$, can be generated with the following parameters: *% attempted shot directives,* $\tilde{s}^T$, and similarly *% successful attempts*, and *% extra successes* (where the user captures more than one shot of footage for a given shot directive); and for each property in every shot directive the hypothetical user's *performance* in following the directive can be indicated as good, average, or poor.

Using synthetic data, $R(\cdot)$'s performance is assessed by observing the change in normalized cumulative difference D between the target tempo, and tempo of the rough cut and repaired movie respectively:

$$D^C = \frac{1}{F}\sum_{f=1}^{F} |T(\tilde{s}^T(f)) - T(\tilde{r}^C(f))| \qquad (4)$$

Where f is frame number, and F the total number of frames in the rough cut. D^R is calculated in a similar manner.

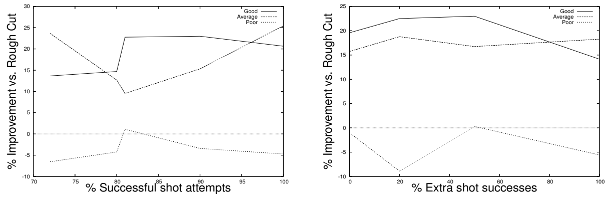


Figure 2: $D^R - D^C$ for various capture parameters.

attempted $\tilde{s}^T$	success	extra success	perf.	D^C	D^R	$D^R - D^C$†
100(%)	100	100	good	.32	.25	.07
100	90	20	good	.32	.24	.08
100	90	0	good	.34	.26	.08
100	100	100	avg.	.42	.31	.11
90	90	20	avg.	.41	.33	.08
100	90	0	avg.	.37	.30	.07
90	90	20	poor	.54	.54	.00

† Figures are averages over 5 synthetic captures. $T(\cdot)$ smoothed $\sigma = 250$.

Table 1: Performance of $R(\cdot)$ for various synthetic captures.

Table 1 contains the performance figures for $R(\cdot)$ across a sample of synthetic capture parameter sets. $D^R - D^C$ is the average improvement of the repaired tempo signal. $R(\cdot)$ maintains an improvement of the order of 20% across a range of parameter sets for capture sessions of *good* or *average* user performance. Surpisingly, $R(\cdot)$ results in a small loss for many *poor* user performance captures, possibly due to over compensation, as tempo difference in Equation 2 is blind to repairs already prepared for preceding shots in r^C, exacerbated by the high impact of motion on tempo, which is greater than shot length in the formulation of Equation 1.

Another surprising outcome is the lack of any consistent trend to $D^R - D^C$ with the availability of extra footage in the form of extra successful captures. Theoretically this should provide $R(\cdot)$ with more opportunity to reduce the deviation from target. However, the potential for improvement is heavily dependent on where missed and additional captures occur. They are placed randomly during synthetic capture generation, and further investigation targetted at this factor will be illuminating. The algorithms of Section 4 run in under 10 seconds on standard desktop hardware.

6. CONCLUSION AND FUTURE WORK

We have demonstrated the automatic repair of visual tempo, an important aesthetic element of video. It is first *injected* by IMCE into a desired set of shots, *detected* in the video resulting from the 'noisy', contingent filming situation, and finally *repaired* to a considerable degree. We plan to attempt automatic repair of other aesthetics, including visual approach, which influences the patterning of framing types in shots, and is an important element for clarifying and intensifying an experience in film.

7. REFERENCES

[1] X.-S. Hua, L. Lu, and H.-J. Zhang, "AVE - Automated home video editing," in *ACM Multimedia*, Nov. 2003, pp. 490–497.

[2] Muvee Technologies, "autoProducer," www.muvee.com.

[3] B. Barry and G. Davenport, "Documenting life: Videography and common sense," in *ICME*, Baltimore, US, July 2003.

[4] M. Davis, "Editing out video editing," in *IEEE Multimedia Magazine, spec. ed. Computational Media Aesthetics*, pp. 54–64. IEEE Computer Society, April-June 2003.

[5] B. Adams, S. Venkatesh, and R. Jain, "IMCE: Integrated media creation environment," *ACM Trans. on Multimedia Computing, Comms. and Applications, To appear*, 2005.

[6] Herbert Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*, 3rd Edition, Wadsworth Pub Company, 1999.

[7] Chitra Dorai and Svetha Venkatesh, "Computational Media Aesthetics: Finding meaning beautiful," *IEEE Multimedia*, vol. 8, no. 4, pp. 10–12, October-December 2001.

[8] B. Adams, C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo," *IEEE Trans. on Multimedia*, vol. 4, no. 4, pp. 472–481, December 2002.

[9] H. Sridharan, H. Sundaram, and T. Rikakis, "Computational models for experiences in the arts and multimedia," in *1st ACM Workshop on Experiential Telepresence, in conjunction with ACM Multimedia*, 2003.

[10] Videomaker(.com) Magazine, "Tips for getting started with video," http://www.videomaker.com/, 2004.