Computational Model of Congruency between Music and Video

1. Our goal

To establish a *computational model* for calculating how much a combination of music and video match well (***congruency***).

× 

Hard rock

Congruency = -0.5

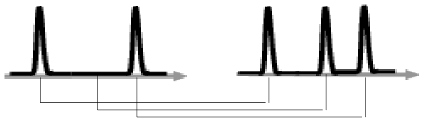
If such a computational model is established…

- A computer system for supporting creating video works can be developed.
  e.g. search for background music that matches the video sources given by the user.
- The model can be a hypothesis of the human mechanism of understanding congruency between music and video.

2. Two types of congruency

Temporal congruency

Synchronization of accents in music and video



[Related work]

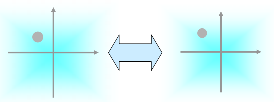
- O. Gillet and G. Rechard: Comparing Audio and Video Segmentations for Music Videos Indexing, Proc. ICASSP 2006.

In the field of psychology, several models have been proposed (e.g. Iwamiya'00).
But they are difficult to implement on a computer.

Semantic congruency

Mood

Similarity of impression people received from music and video



No computational models for mood congruency have been proposed.

Symbolic semantics

e.g. Use of a particular song in a shop means that the time of the shop closing is approaching.

Culture-dependent

We left this as future work.

3. Our computational model

Temporal congruency Correlation between accent feature serieses extracted from music and video

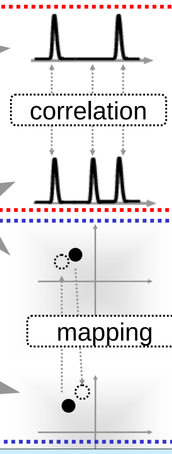
musical signal

- onset sequence
- intensity feature
- timbre feature

feature extraction

video signal

- motion sequence
- shot sequence
- color feature
- motion feature
- shot feature



temporal congruency → congruency ← semantic congruency

$$Cor(\mathrm{M}||\mathrm{V}) = \frac{\sum_t f'_{m,t} f'_{v,t}}{\sqrt{\sum_t (f'_{m,t})^2 \sum_t (f'_{v,t})^2}}$$

Musical accent feature series; Visual accent feature series

$$Con(\mathrm{M}||\mathrm{V}) = \underbrace{(1-\alpha)Cor(\mathrm{M}||\mathrm{V})}_{\text{Temporal congruency}} + \underbrace{\alpha Dist(\mathrm{M}||\mathrm{V})}_{\text{Semantic congruency}}$$

$$Dist(\mathrm{M}||\mathrm{V}) = \{d(x_m, \hat{x}_m) + d(x_v, \hat{x}_v)\}/2$$

Original music features; Associated music features; Original video features; Associated video features

Semantic congruency

Mutual mapping between musical and visual mood spaces

Key idea Transforming a musical mood vector to a visual mood space
= Associating a video that matches the given music

If the video associated from the music is similar to the original video, the music and video should match.

The 10th Int'l Conf. on Music Perception and Cognition (24-30th Oct. 2008) in Sapporo, Japan

Tetsuro Kitahara[*1,*3], *Masahiro Nishiyama*[*2], *and Hiroshi G. Okuno*[*2,*3]

[*1]*Kwansei Gakuin University, Japan* [*2]*Kyoto University, Japan*

[*3]*CrestMuse Project, CREST, JST, Japan*

t.kitahara@kwansei.ac.jp, http://ist.ksc.kwansei.ac.jp/~kitahara/

4. Details of semantic (mood) congruency calculation

[Our strategy] Mutual mapping between musical and visual mood spaces

[One possible solution] Transform musical and visual feature spaces to a common mood space consisting of several adjectives

→ We do not want to do it because adjectives are not sufficient to represent musical and visual mood

[Our solution] Directly transform musical and visual mood spaces by a linear transformation that can be trained with only pairs of congruent music and videos (No teacher signals of mood adjectives are needed.)

[Method]

1. Musical feature vector g'_{m} and visual feature vector g'_{v} are transformed to new vectors x_{m} and x_{v} using PCA.
$$g'_{\mathrm{m}} \simeq A_{\mathrm{m}} x_{\mathrm{m}}, \quad g'_{\mathrm{v}} \simeq A_{\mathrm{v}} x_{\mathrm{v}}$$
2. The concatenated vector of x_{m} and x_{v} are transformed into a new vector c using PCA again.
$$x = \begin{pmatrix} x_{\mathrm{m}} \\ x_{\mathrm{v}} \end{pmatrix} \simeq Pc = \begin{pmatrix} P_{\mathrm{m}} \\ P_{\mathrm{v}} \end{pmatrix} c$$
When appropriately congruent pairs of music and video are given, the generated space can be considered an integrated mood space.
3. After the given musical and visual feature vectors are transformed to this integrated mood space, they are transformed to visual and musical feature space.
$$\hat{x}_{\mathrm{v}} = P_{\mathrm{v}} P_{\mathrm{m}}^{-} x_{\mathrm{m}}, \quad \hat{x}_{\mathrm{m}} = P_{\mathrm{m}} P_{\mathrm{v}}^{-} x_{\mathrm{v}}$$
4. The similarity between the original and transformed features is calculated using the cosine distances.
$$Dist(\mathrm{M}||\mathrm{V}) = \{d(x_{\mathrm{m}}, \hat{x}_{\mathrm{m}}) + d(x_{\mathrm{v}}, \hat{x}_{\mathrm{v}})\}/2$$

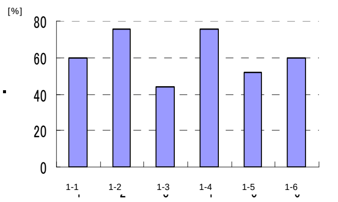
5. Experiments

Experiment I: Temporal congruency

[Our hypothesis] Temporal congruency is more important for music-oriented works.

- 5 subjects rated congruency for every combination.
- Our model's and subjects' results were binarized and accuracy rates were calculated.
- Average accuracy: 61.3%

Music-oriented works (from "Fantasia")
(1-1) Symphony No. 5 (Beethoven)
(1-2) Pines of Rome
(1-3) Rhapsody in Blue
(1-4) Piano Concerto No. 2
(1-5) The Sorcerer's Apprentice
(1-6) Pomp and Circumstance Marches

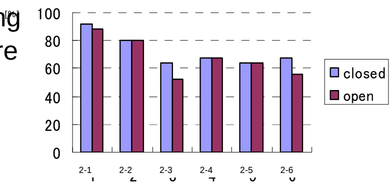


Experiment II: Semantic congruency

[Our hypothesis] Semantic congruency is more important for video-oriented works.

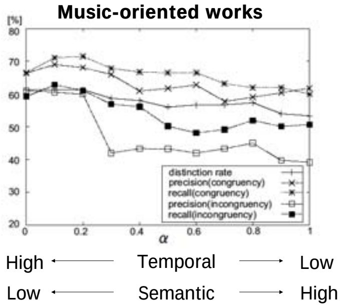
- Same experimental conditions as Experiment I.
- Both closed and open training of mood space mapping were tried.
- Average accuracy: 68.0%

Video-oriented works
(2-1) Pirates of the Caribbean
(2-2) Star Wars Episode I
(2-3) Star Wars Episode II
(2-4) Catch Me If You Can
(2-5) Back to the Future
(2-6) The Phantom of the Opera

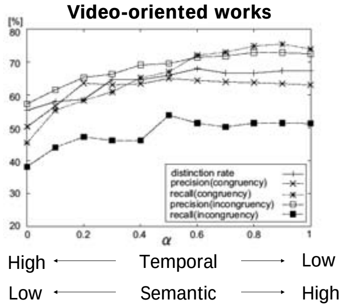


Experiment III: Integration with different weights

Music-oriented works

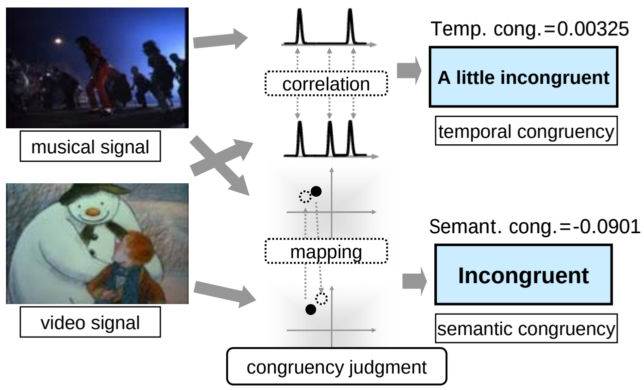


Video-oriented works



→ Our hypotheses were supported.

Example of congruency calculation



Appendix: Features used

I. Musical features

Intensity	Sum of intensities for all frequency bins
Sub-band intensity	Intensity of each sub-band (7 sub-bands prepared)
Spectral centroid	Centroid of the short-time amplitude spectrum
Spectral rolloff	85th percentile of the spectral spectrum
Spectral flux	2-norm distance of the frame-to-frame spectral amplitude difference
Bandwidth	Amplitude weighted average of the differences between the spectral components and the centroid
Sub-band peak	Average of the percent of the largest amplitude values in the spectrum of each sub-band
Sub-band valley	Average of the percent of the lowest amplitude values in the spectrum of each sub-band
Sub-band contrast	Difference between "peak" and "valley" in each sub-band

II. Visual features

Mean and var. of L-values in CIELUV color space
Color histogram in CIELUV color space
Mean and var. of Y-, U-, and V-values in YUV color space
Temporal differential of optical flow
Temporal differential of color histogram in CIELUV color space
Temporal differential of color histogram in YUV color space