6 Seconds of Sound and Vision: Creativity in Micro-Videos

Miriam Redi[1] Neil O'Hare[1] Rossano Schifanella[3,*] Michele Trevisiol[2,1] Alejandro Jaimes[1]

[1]Yahoo Labs, Barcelona, Spain {redi,nohare,ajaimes}@yahoo-inc.com
[2]Universitat Pompeu Fabra, Barcelona, Spain {trevisiol}@acm.org
[3]Università degli Studi di Torino, Torino, Italy {schifane}@di.unito.it

Abstract

The notion of creativity, as opposed to related concepts such as beauty or interestingness, has not been studied from the perspective of automatic analysis of multimedia content. Meanwhile, short online videos shared on social media platforms, or micro-videos, have arisen as a new medium for creative expression. In this paper we study creative micro-videos in an effort to understand the features that make a video creative, and to address the problem of automatic detection of creative content. Defining creative videos as those that are novel and have aesthetic value, we conduct a crowdsourcing experiment to create a dataset of over 3,800 micro-videos labelled as creative and non-creative. We propose a set of computational features that we map to the components of our definition of creativity, and conduct an analysis to determine which of these features correlate most with creative video. Finally, we evaluate a supervised approach to automatically detect creative video, with promising results, showing that it is necessary to model both aesthetic value and novelty to achieve optimal classification accuracy.

1. Introduction

Short online videos, or *micro-videos*, have recently emerged as a new form of user-generated content on social media platforms such as Vine, Instagram, and Facebook[1]. The Vine platform, in particular, has become associated with the notion of creativity, as it was launched with the goal of allowing users to create 6-second videos whose time constraint "inspires creativity"[2]. Some commentators have even claimed of Vine in particular that "its constraints were allowing digital videos to take on entirely new forms"[3], and interest in Vine videos has prompted the creation of a specific 6-second film category at major film festivals such as the Tribeca Film Festival in New York.

Not all micro-videos uploaded on social media platforms are creative in nature (1.9% of randomly sampled videos were annotated as creative in our study), and quality can vary widely. This motivates the need for automatic approaches to detect and rank the best, and in particular the most *creative*, micro-video content on social media platforms. Such applications can increase the visibility of video authors, and replace or augment current features of social-media platforms such as "Editors Picks", which showcases the best content on Vine.

Micro-videos provide a unique opportunity to address the study of audio-visual creativity using computer vision and audio analysis techniques. The very short nature of these videos means that we can analyze them at a micro-level. Unlike short video sequences within longer videos, the information required to understand a micro-video is contained within the video itself. This allows us to study audio-visual creativity at a fine-grained level, helping us to understand what, exactly, constitutes creativity in micro-videos.

In this paper we study the audio-visual features of *creative* vs *non-creative* videos[4] and present a computational framework to automatically classify these categories. In particular, we conduct a crowdsourcing experiment to annotate over 3,800 Vine videos, using as guidelines: (1) a widely accepted definition of creative artifacts as those that are *novel* and *valuable*, and (2) insights from the philosophy of aesthetics about the judgements of aesthetic value (i.e. sensory, emotional/affective, and intellectual). We go on to use this dataset to study creative *micro-videos* and to evaluate approaches to automatic detection of creative micro-videos.

The main contributions of this paper are:

- We create a new dataset of creative *micro-videos*, and make the vine video ids and annotations publicly available to the research community[5].

*This work has been performed when the author was a Visiting Scientist at Yahoo Labs, Barcelona, within the framework of the FREP grant.

[1] http://vine.co, http://instagram.com, http://facebook.com

[2] http://blog.vine.co/post/55514427556/introducing-vine

[3] https://medium.com/art-technology/a4433fb334f

[4] Throughout the paper we will use the word "video" to refer to "micro-videos" of a few seconds

[5] available for download at: http://di.unito.it/vinecvpr14

- We propose and implement a new set of features to model the novelty and aesthetic value of *micro-videos*.
- We analyze the extent to which each of these features, and other existing features, correlate with creativity, giving insights into the audio-visual features most associated with creative video. We also classify videos as creative/non-creative, with promising results, and we show that combining aesthetic value and novelty features gives highest accuracy.

Unlike previous work in computational aesthetics [5, 7], which mainly focuses on assessing visual beauty using compositional features, we explore here the more complex and subtle concept of *creativity*. Focusing on creative content allows us to analyze audio-visual content from a different perspective, allowing us to model the fact that creative content is not always the most beautiful-looking (in the conventional sense) or visually interesting. To the best of our knowledge, this is the first work to address creativity in micro-videos.

In the next Section we present related work, and we define video creativity in Section 3. In Section 4 we describe a crowdsourced annotation of Vine videos. Section 5 presents computational features for modeling creativity. In Section 6 we correlate these features with, and evaluate the automatic classification of, creative content. We conclude in Section 7.

2. Related Work

Our work is closely related to computational approaches to studying concepts such as beauty [5], interestingness [7], memorability [10], or emotions [17]. In particular, we are influenced by recent work in computational aesthetics for the automatic assessment of visual beauty. The earliest work [5, 12] distinguishes between high-quality (professional) and low-quality (amateur) photos based on features inspired by photographic rules, with applications in image quality enhancement [3] and automatic aesthetic feedback for photographers [32]. Nishiyama *et al*. [25] propose more complex visual features based on color harmony, and combine them with low-level features for aesthetic image classification. Other work has investigated generic local features for modeling beauty, showing that they outperform systems based on compositional features [19]. Several researchers have included the semantics of the image in the aesthetic evaluation, labeling images according to their scene content and building category-based beauty models [16, 23].

The main difference between visual aesthetic research and our work is that the notion of *creativity* is more complex than visual photographic beauty, in addition to the fact that we also focus on audio. We argue that creative videos may not be always considered 'beautiful' in the conventional sense, and may even be 'ugly'. While we incorporate and re-elaborate many of the mentioned approaches for detecting creative videos, by using sensory (including aesthetic), and visual affect features, we also design a new set of features to model audio-visual creativity.

Moreover, while much related work focuses on still images, in our work we analzye *video* data, and we build specific video features for micro-videos. The few previous works on video aesthetics build video features based on professional movie aesthetics [4, 2], or simply aggregate frame-level features [21], with limited success.

Also different from much of the work in computational aesthetics, we use a croudsourced groundtruth, allowing us to create a high quality labelled dataset using a set of annotation guidelines tailored for creativity. Crowdsourcing was previously used to build a corpus for image memorability [10], but most computational aesthetics research exploits online professional photo websites such as *dpchallenge.com* [5, 12, 7, 23, 16], *photo.net* [16], or Flickr [7].

3. Defining Video Creativity

Although the precise definition of creativity has been the subject of debate in many disciplines, one of the most common observations is that creativity is connected with imagination and innovation, and with the production of novel, unexpected solutions to problems [24]. However, *"All who study creativity agree that for something to be creative, it is not enough for it to be novel: it must have value, or be appropriate to the cognitive demands of the situation"* [31], an idea that is shared by many researchers [8, 22, 31]. Based on these observations, we define a creative artifact as one that is *novel* (surprising, unexpected) and has *value*.

As applied to micro-videos, we interpret by *novelty* that the video is unique in a significant way, or that it expresses ideas in an unexpected or surprising manner. *Value* is a more complex notion, however, and in this context it is best equated with aesthetic value. Most definitions of aesthetic value incorporate the maxim that beauty is in the eye of the beholder: Immanuel Kant, for example, in his *Critique of Judgement*[11], argues that aesthetic judgements involve an emotional response (*e.g*., pleasure) to a sensory input (i.e. the audio-visual signal from the video) that also provokes "reflective contemplation". At the risk of oversimplifying, judgements of aesthetic value involve *sensory*, *emotional* and *intellectual* components.

In the following sections, we will use this definition to: (1) provide a definition of creative video as part of our guidelines for crowd workers to annotate videos as creative or non-creative (Section 4), and (2) inform our choice of computational features for modeling creative videos.

4. Dataset

To create a corpus of micro-videos annotated as *creative*, we first identified a set of candidate videos that were likely to be creative. This was necessary because our preliminary

analysis showed that only a small fraction of videos are creative, meaning that a random sampling would need an extremely large annotation effort to collect a reasonable number of positive creative videos to analyze. With this in mind, we defined a set of sampling criteria likely to return creative videos. We started by sampling 4,000 videos. Specifically, we took (a) $1,000$ videos annotated with hashtags that were associated to creative content by 3 different blogs about Vine: *#vineart, #vineartist, #artwork, and #vineartgallery* (b) 200 videos *mentioned* in 16 articles about Vine creativity on social media websites, (c) $2,300$ videos authored by the 109 creators of the videos identified in criteria *b*, based on the assumption that these authors are likely to author other creative micro-videos, and (d) 500 randomly selected videos from the Vine streamline, for the purpose of estimating the true proportion creative videos on Vine. The results of the labeling experiment summarized in Table 3 confirm the validity of this sampling strategy: while only 1.9% of the random sample has been labeled as creative (D-100), our sampling strategy yielded 25% creative videos, giving a corpus that is large enough to be useful. In total, after discarding invalid urls, we annotated 3,849 candidate videos, created and shared between November 2012 to July 2013.

We annotate these videos using Crowdflower[6], a large crowdsourcing platform. To ensure quality annotations, the platform enables the definition of *Gold Standard* data where workers are assigned a subset of pre-labelled 'jobs', allowing the known true label to be compared against the contributor label. This mechanism allows worker performance to be tracked, and can ensure that only judgements coming from competent contributors are considered. It also presents an opportunity to give feedback to workers on how to improve their annotations in response to incorrect answers.

In the experiment, a contributor looks at a 6-second video and judges if it is creative. According to Section 3, a creative video is defined as a video that: (1) has aesthetic value, or evokes an emotion (happy, sad, angry, funny, etc), and (2) has interesting or original/surprising video/audio technique. The worker is advised to listen to the audio, and can watch a pair of exemplar *creative* and *non-creative* videos before performing the job. After watching the target video the contributor answers the question *"Is this video creative?"* with "positive", "negative" or "don't know". In the first two cases, the user can give more details of the motivation of their choice according to the criteria in Table 1, phrased in a simple language appropriate to crowdsourcing platforms, where workers typically do not take time to read complex definitions and guidelines [20]. To ensure that the job could be easily understood by crowd workers, in a preliminary survey we collected feedback on the interface from 15 volunteers.

The experiment ran for 5 days and involved 285 active workers (65 additional workers were discarded due to the low quality of their annotations) located in USA (88%), United Kingdom (8%), and Germany (4%).[7] No time constraint was set on the task, and each video was labeled by 5 independent workers. The final annotations reached a level of 84% worker agreement (82% for creative, 85% for non-creative), which we consider high for this subjective task. Looking at per-video agreement, summarized in Table 2, 48% of videos have 100% agreement (i.e. all 5 independent annotators agreed), 77% show an 80% consensus. These levels of agreement represent different criteria for labeling a video as (non) creative, and in Section 6 we consider 3 different labelled ground-truth datasets, D-100, D-80, and D-60, based on 100%, 80% and 60% agreement. From Table 2 we can also see that 25-30% of videos were annotated as creative.

Aesthetic Value	*Sensory*	The audio is appealing/striking The visuals are appealing/striking
	Emotional	The video evokes an emotion
	Intellectual	The video suggests interesting ideas
Novelty	The audio is original/surprising The visuals are original/surprising The story or content is original/surprising	

Table 1. Criteria for labeling a video as creative

Dataset	% Videos	# Creative (%)	# Non-creative (%)
D-60	100%	1141 (30%)	2708 (70%)
D-80	77%	789 (27%)	2196 (73%)
D-100	48%	471 (25%)	1382 (75%)

Table 2. Summary of the results of the labeling experiment. D-60: videos with at least 60% agreement between annotators. D-80: at least 80% agreement. D-100: 100% agreement.

	(a) Hashtags	(b) Blogs	(c) Creators	(d) Random
Creative	34.05%	79.57%	27.41%	1.88%
Non-Creative	65.95%	20.43%	72.59%	98.12%

Table 3. Creative vs non-creative videos per sampling strategy, for the D-100 dataset (100% agreement).

Table 3 shows the distribution of creative and non-creative videos according to the strategy used to sample the videos. As expected, the videos specifically mentioned in blogs about Vine (b) have the highest proportion of creative videos, while the vast majority of randomly sampled videos (d) are non-creative, justifying the need for our sampling strategies.

5. Features for Modeling Creativity

In this Section we describe novel and existing features for modeling creative micro-videos, which we group based on the two components of our definition of creative videos: novelty and value. We re-use existing features from computational aesthetics, semantic image analysis, affective image classification, and audio emotions modeling, and propose

[6] http://www.crowdflower.com

[7] Additional demographic information was not available.

Group	Feature	Dim	Description
AESTHETIC VALUE			
Sensory Features			
Scene Content	*Saliency Moments* [26]	462	Frame content is represented by summarizing the shape of the salient region
Filmmaking Technique	*General Video Properties*	2	*Number of Shots*, *Number of Frames*
	Stop Motion	1	Number of non-equal adjacent frames
	Loop	1	Distance between last and first frame
	Movement	1	Avg. distance between spectral residual [9] saliency maps of adjacent frames
	Camera Shake	1	Avg. amount of camera shake [1] per frame
Composition and Photographic Technique	*Rule of Thirds* [5]	3	HSV average value of the inner quadrant of the frame (*H(RoT),S(RoT),V(RoT)*)
	Low Depth of Field [5]	9	LDOF indicators computed using wavelet coefficients
	Contrast [6]	1	Ratio between the sum of max and min luminance values and their difference
	Symmetry [27]	1	Difference between edge histograms of left and right halves of the image
	Uniqueness [27]	1	Distance between the frame spectrum and the average image spectrum
	Image Order [28]	2	Order values obtained through Kologomorov *Complexity* and Shannon's Entropy
Emotional Affect Features			
Visual Affect	*Color Names* [17]	9	Amount of color clusters such as red, blue, green, ...
	Graylevel Contrast Matrix Properties [17]	10	*Entropy, Dissimilarity, Energy, Homogeneity* and *Contrast* of the GLCM matrix
	HSV statistics [17]	3	Average *Hue, Saturation and Brightness* in the frame
	Pleasure, Arousal, Dominance [30]	3	Affective dimensions computed by mapping HSV values
Audio Affect	*Loudness* [15]	2	Overall *Energy* of signal and avg *Short-Time Energy* in a 2-seconds window
	Mode [15]	1	Sums of key strength differences between major keys and their relative minor keys
	Roughness [15]	1	Avg of the dissonance values between all pairs of peak in the sound track spectrum
	Rythmical Features [15]	2	*Onset Rate* and *Zero-Crossing Rate*
NOVELTY			
Novelty	*Audio Novelty*	10	Distance between the audio features and the audio space
	Visual Novelty	40	Distance between the visual features and each visual feature space

Table 4. Audiovisual features for creativity modeling

new features to represent filmmaking technique and novelty. Table 4 summarizes all the features introduced in this section.

5.1. Aesthetic Value Features

We use a set of features to model the aesthetic value of a video based on two of the three components of aesthetic value identified in Section 3: the *sensory* component and the *emotional* affect of the video. The third, *intellectual*, component is, to the best of our knowledge, not modeled by any existing computational approaches, so we do not model it in this work.

5.1.1 Sensory Features

Sensory features model the raw sensory input perceived by the viewer, which can be approximated by the raw signal output by the video. Such features cover all aspects of the signal, i.e. visual, audio, movement, filmmaking techniques, etc. We implement existing features for semantic image classification and aesthetic image analysis, and we design new descriptors to capture the structural characteristics of short-length online videos.

Video Scene Content. We extract the 462-dimensional namely the Saliency Moments feature [26] from video frames, a holistic representation of the content of an image scene based on the shape of the salient region, which has proven to be extremely effective for semantic image categorization and retrieval.

Composition and Photographic Technique. In computational aesthetics, several compositional descriptors describing the photographic and structural properties of images and video frames have been proposed. Other features attempt to model the visual theme of images and videos [29]. We use some of the most effective frame-level compositional features, such as the *Rule of Thirds* and *Low Depth of Field* [5], the *Michelson Contras*t [6], a measure of *Symmetry* [27], and a *Uniqueness* [27] measure indicating the familiarity of the spatial arrangement. Finally we implement a feature describing the *Image Order* using information theory-based measurements [28].

Filmmaking Technique Features. We design a set of new features for video motion analysis, inspired by movie theory and tailored to model the videomaking techniques of short on-line videos.

General Video Properties. We compute the number of frames N_f and the number of shots N_s in the video. In the current setting, the number of frames is a proxy for frame rate, as almost all videos are exactly 6 seconds in length, whereas the frame rate tends to vary.

Stop Motion. Many popular creative short videos are stop-motion creations, where individual photos are concatenated to create the illusion of motion. In such videos the frequency of changes in the scene is lower than traditional videos. We capture this technique by computing the Euclidean distance $\delta(F_i, F_{i+1})$ between the pixels of neighboring frames F_i

and F_{i+1} and then retaining as a stop motion measure S the ratio between N_f and the number of times such difference is not null (the scene is changing), namely

$$S = \frac{N_f}{1 + \sum_{i=1}^{N_f - 1} sgn(\delta(F_i, F_i + 1))}. \quad (1)$$

Loop. Many popular videos in Vine are shared with the hashtag #loop. A looping video carries a repeatable structure that can be watched repeatedly without perceiving where the beginning/end of the sequence is. To capture this, we compute the distance between the first and the last frames of the video, namely $L = \delta(F_1, F_{N_f})$

Movement. similar to previous works, [4, 2], we compute the amount of motion in a video using a feature that can describe the speed of the main objects in the image regardless of their size. We first compute a saliency map of each frame and then retain, as a movement feature, the average of the distances between the maps of neighboring frames:

$$M = 1/N_f \sum_{i=1}^{N_f - 1} \delta(SM(F_i), SM(F_{i+1})) \quad (2)$$

where $SM(\cdot)$ is the saliency map computed on the frame using the Spectral Residual technique [9].

Camera-Shake. Typical micro-videos are not professional movies, and often contain camera shake introduced by hand-held mobile phone cameras. Artistic video creators, however, often carefully produce their videos, avoiding camera-shake. We compute the average amount of camera shake in each frame using an approach based on the directionality of the Hough transform computed on image blocks [1].

5.1.2 Emotional Affect Features

In this section we separately introduce sets of visual and audio features known to correlate with emotional affect.

Visual Affect. We extract a set of frame level affective features, as implemented by Machajdik & Hanbury [17], namely *Color names*, *Graylevel Contrast Matrix (GLCM) properties*, *Hue, Saturation and Brightness* statistics, *Level of Detail*, and the *Pleasure*, *Arousal*, and *Dominance* values computed from HSV values [30] .

Audio Affect. Inspired by Laurier et al [15], we implement, using the MIRToolbox [14], a number features for describing audio emotions, collecting them a 6-dimensional feature vector. We describe the sound *Loudness*, the overall volume of the sound track, its *Mode* (indicating if the sound in the Major or Minor mode), the audio *Roughness* (dissonance in the sound track), and *Rythmical Features* describing abrupt rhythmical changes in the audio signal.

5.2. Novelty

The novelty of an artifact can be represented by its distance from a set of other artifacts of the same type. One way to compute such distance is to first divide the attribute space into K clusters, and then calculate the distance the between the artifact and its nearest cluster [18]. In our approach, we compute an improved novelty feature that takes into account the distances between the artifact attribute and *all the clusters* in the attribute space, thus measuring not only the distance to the most similar element, but the detailed position of the attribute in the space.

We measure novelty for both the visual and the audio channel of the video, using as attributes the aesthetic values features from Section 5.1. We take a random set of videos, independent of our annotated corpus, and extract the 4 groups of visual attributes (*Scene Content (SC)*, *Filmmaking Techniques*, *Composition and Photographic Technique* and *Visual Affect*), and the *Audio Affect* attributes. We cluster the space resulting from each attribute into 10 clusters using K-means, obtaining 40 clusters for the visual attributes (10 clusters each for 4 attributes) and 10 for the audio attribute.

To calculate the novelty score for a given video, we extract the visual and audio attributes, and we then compute the **Audio Novelty** as the collection of the distances between the *Audio Affect* attribute of the video and all the clusters of the corresponding space (giving a 10 dimensional feature). Similarly, we compute the video **Visual Novelty** as the set of distances between each visual attribute of the video and the corresponding cluster set (40 dimensions).

6. Experimental Results

In this Section we explore the extent to which audio-visual features correlate with creative video content, and then evaluate the approaches for creative video classification.

6.1. What Makes a Video Creative?

To analyze which features correlate most with creative micro-videos, we consider videos with 100% agreement (i.e. D-100 from Table 2), as we are interested in the correlations for the cleanest version of our dataset. We extract 7 groups of features for each video: *Scene Content, Composition/Photographic Technique, Filmmaking Technique, Visual Emotional Affect, Audio Emotional Affect, Visual Novelty,* and *Audio Novelty.* For frame-level features, we consider the features of middle frame of the video.

We first analyze to what extent each group of features correlates with video creativity, using the *Multiple Correlation Coefficient* (MPC), which measures how well a multidimensional variable fits a monodimensional target variable, given the reconstructed signal after regression. In our context, the elements of the multidimensional variable are the individual features within a feature group.

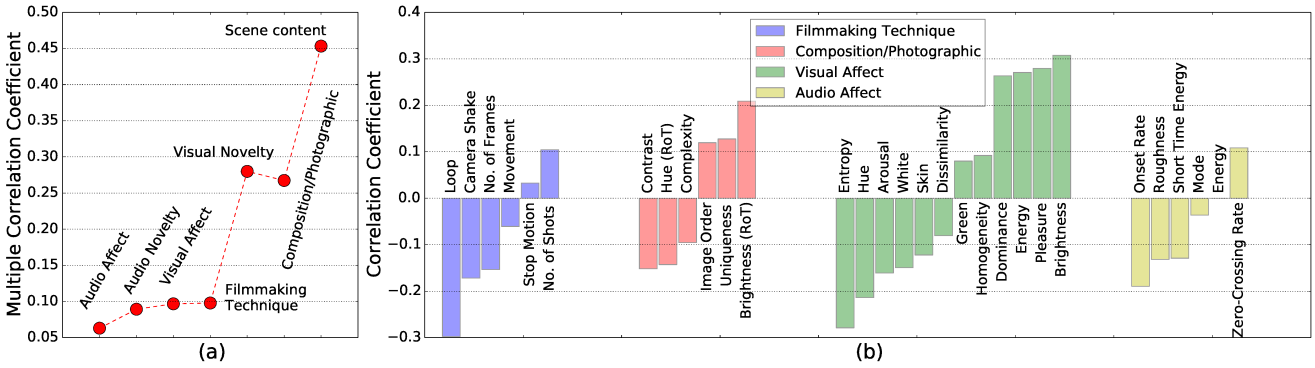

Figure 1. Analysis of the most relevant features and components for video creativity prediction

In Figure 1(a), we report MPC values for all features, testing how well each of our groups of features fits the label vector of our data. The results show that *Scene Content* is most strongly correlated with creative videos, followed by *Composition/Photographic* features and *Video Novelty*, showing that both novelty and aesthetic value features are crucial for understanding creativity. Emotional affect features are less strongly correlated than sensory features, suggesting that the raw sensory output of the audiovisual stream is more useful than features that attempt to model emotional affect. The correlation for audio features is much lower than for visual features: it seems likely that micro-video authors and annotators both place more emphasis on the visual aspect.

To determine, within each group of features, the most important individual features, we calculate the Pearson correlation coefficient ρ between individual features and creative labels. For this analysis, we exclude the *Scene Content* features, whose elements are non-separable, and the *Novelty* features, which require all features to represent the position of the video in the attribute space.

In Figure 1(b) we show the ρ values for the most highly correlated features. Among the *Composition/Photographic Technique* and *Visual Affect* features, we can see that creative videos strongly correlate with visual uniformity (positive correlation with *Energy*), and order (negative correlation with *Entropy* and the *Complexity* measure), suggesting that videos with homogenous frames are more likely to be perceived as creative. Moreover, a negative correlation with the *Hue* statistic in both the whole image (*Visual Affect - Hue*) and in the inner quadrant (*Composition/Photographic - Hue (RoT)*) and a highly positive ρ for *Visual Affect - Brightness* and *Composition/Photographic - Brightness (RoT)* shows that creative videos are related to warm and bright colors. The *Uniqueness* feature shows a positive ρ, indicating videos with a less familiar layout are more likely to be labeled as creative. Surprisingly, important properties for visual aesthetics such as *Symmetry* ($\rho = 0.04$) and *Low Depth of Field* ($\rho = 0.02$) do not play a key role for creative video detection (these are not shown in Figure 1(b), which only includes features with a high correlation). Also, *Skin* color (Visual Affect) is negatively correlated with creativity, showing that the presence of people is not so common in creative videos, unlike most other popular videos.

Creative videos are associated with highly pleasant emotions, and dominant, non-overwhelming, controllable emotions ($\rho = 0.24$ for *Pleasure* and $\rho = 0.23$ for *Dominance*). In terms of *Audio Affect*, *Onset Rate* and loudness (*Short Time Energy*) are negatively correlated with creativity, meaning that less frenetic, low-volume sound is preferred.

Regarding *Filmmaking Techniques*, the most highly correlated feature is the presence of the loop technique, a common form of expression in micro-videos (it is negatively correlated because a high score indicates low likelihood of a loop). *Camera Shake*, which can be seen as an inverse quality measure, is a negative indicator of creativity, suggesting that more 'polished' videos are likely to be seen as creative.

As we can see from our findings, creativity involves a variety of dimensions beyond photographic beauty. Photographic features traditionally used in visual aesthetic frameworks [5, 7] are equally or even less correlated with creativity than other, complementary, features. Moreover, we can see that single *Filmmaking Technique* features such as *Loop*, or *Audio Affect* features such as *Onset Rate*, are better indicators of creativity, compared to single photographic features.

6.2. Classifying Creative Videos

We now evaluate methods to automatic classify videos as creative, using the same features.

Experimental Setup. In Table 2 in Section 4 we described three different versions of our annotated corpus, based on 60%, 80% and 100% per-video annotation agreement. These datasets give a natural tradeoff between dataset size and label accuracy. To measure the effect of this tradeoff on classification, we report results for each dataset.

For each version of the corpus we use 2/3 of the positive examples for training, and the rest for testing. For training and testing, we subsample an equal amount of negative ex-

amples, to ensure a balanced set. We train a separate Support Vector Machine with Radial Basis Function (RBF) kernel for each of the 7 groups of features. For groups of features that are calculated for a single video frame, at the training stage we sample 12 frames for the video, and create a separate training instance for each sampled frame, each given the label of the parent video. We use the trained models to classify the creative videos in the test set. For each video, the classifier outputs a label and a classification score. For the frame-level features, we sample 12 frames as in training, classify each, and retain as overall classification of the video the rounded average of the single frame scores. We use classification accuracy as our evaluation measure.

For the novelty features, we use 1000 non-annotated videos for the clustering. To check that this number does not introduce any bias in our experiment, we re-compute clustering on an increasing number of videos, from 500 to 5000, and obtained similar results as those presented in Table 5.

To test the complementarity of the groups of features and the improvement obtained by combining them, we also combine the classification scores of different classifiers using the *median* value of the scores of all the classifiers, previously shown to perform well for score aggregation [13].

Results. The classification results are shown in Table 5. Similar to the correlations, we can see that the best feature group is *Composition/Photographic Technique*, with 77% accuracy (D-100 dataset), followed by *Scene Content* and *Filmmaking Technique* features. We can also see that Emotional Affect features are outperformed by Sensory features. Our new, 6-dimensional, *Video Technique* feature achieves comparable classification accuracy to the performance of the 462 dimension *Scene content* feature. Combining emotional and sensory features improves classification accuracy to 79%, showing the complementarity of these features.

Feature	**Accuracy**		
	D-60	**D-80**	**D-100**
Aesthetic Value			
Sensory Features			
Scene Content	0.67	0.69	0.74
Filmmaking Techniques	0.65	0.69	0.73
Composition & Photographic Technique	0.67	0.74	**0.77**
All Sensory Features	0.69	**0.75**	0.77
Emotional Affect Features			
Audio Affect	0.59	0.53	0.67
Visual Affect	0.65	0.66	0.66
All Emotional Affect Features	0.62	0.56	**0.71**
All Aesthetic Value Features	0.68	0.72	**0.79**
Novelty			
Audio	0.58	0.58	0.63
Visual	0.63	0.67	**0.74**
Audio + Visual Novelty	0.59	0.63	0.69
Novelty + Aesthetic Value	0.69	0.73	**0.80**

Table 5. Prediction results for value and novelty features

Although the *Novelty* features carry some discriminative power for creative video classification, Aesthetic Value features are still more discriminative. However, when we combine novelty and value features, we can see their complementarity, with the classification accuracy increased from 79% to 80% for the D-100 dataset.

Overall, we can notice the importance of using a diversity of features for creativity prediction, since classifiers based on traditional photographic features or generic scene features, typical of visual aesthetic frameworks, benefit from the combination with other cues, justifying a tailored framework for creative video classification.

Finally, we can also see that the quality of the annotations is crucial: classification accuracy is always much higher for the cleanest dataset, D-100, even though this dataset is only 60% the size of the D-80 dataset, and less than half the size of the D-60 dataset.

7. Conclusions

In this paper, we study creativity in short videos, or *micro-videos*, shared in online social media platforms such as Vine or Instagram. Defining creative videos as videos that are *novel* (*i.e.*, surprising, unexpected) and have *aesthetic value*, we run a crowdsourcing experiment to label more than 3,800 *micro-videos* as creative or non-creative. We obtain a high level of inter-annotator agreement, showing that, with appropriate guidelines, it is possible to collect reliable annotations for a subjective task such as this. From this annotation we see that a small, but not insignificant, 1.9% of randomly sampled videos are labeled as creative.

We propose a number of new and existing computational features, based on *aesthetic value* and *novelty*, for modeling creative *micro-videos*. We show that groups of features based on scene content, video novelty, and composition and photographic technique are most correlated with creative content. We show that specific features measuring order or uniformity correlate with creative videos, and that creative videos tend to have warmer, brighter colors, and less frenetic, low volume sounds. Also, they tend to be associated with pleasant emotions, and dominant, non-overwhelming, controllable emotions. Loop and Camera Shake features, specifically designed for modeling creativity in *micro-videos*, also show high correlation with creativity. Several features traditionally associated with beauty or interestingness show low correlations with creative *micro-video*, underlining the difference between creativity and those concepts. Specifically, skin color, symmetry and low depth, which are widely used in modeling beauty and interestingness, are not correlated with creative *micro-videos*.

Finally, we evaluate approaches to the automatic classification of creative *micro-videos*. We show promising results overall, with a highest accuracy of 80% on a balanced dataset. The best results are achieved when we combine novelty fea-

tures with aesthetic value features, showing the usefulness of this twofold definition of creativity. We also show that high quality ground truth labels are essential to train reliable models of creative micro-videos.

In future work, we plan to enlarge the set of features for modeling creativity. We will design features to model the intellectual aspect of aesthetic value through semantic visual cues such as specific visual concept detectors. Moreover, we plan to include non-audiovisual cues such as the metadata related to the video (tags, tweets, user profile), the comments about it, and its' popularity in the social media community.

Furthermore, we would like to apply our model, or a modified version of it, to other micro-video platforms and also to a broader spectrum of multimedia content, such as images and longer videos, *etc*., and to study the differences and commonalities between their creative features.

References

[1] http://www.cs.bgu.ac.il/ ben-shahar/teaching/computational-vision/studentprojects/icbv121/icbv-2012-1-kerendamari-bensimandoyev/index.php.

[2] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah. Towards a comprehensive computational model foraesthetic assessment of videos. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 361–364. ACM, 2013.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM Multimedia*, pages 271–280, 2010.

[4] S. Chung, J. Sammartino, J. Bai, and B. A. Barsky. Can motion features inform video aesthetic preferences? Technical Report UCB/EECS-2012-172, EECS Department, University of California, Berkeley, Jun 2012.

[5] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *IEEE ECCV*, pages 288–301, 2006.

[6] M. Desnoyer and D. Wettergreen. Aesthetic image classification for autonomous agents. In *ICPR*, 2010.

[7] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE CVPR*, pages 1657–1664.

[8] L. F. Higgins. Applying principles of creativity management to marketing research efforts in high-technology markets. *Industrial Marketing Management*, 28(3):305–317, 1999.

[9] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE CVPR*, pages 1–8, 2007.

[10] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE CVPR*, pages 145–152. ACM, 2011.

[11] I. Kant and W. S. Pluhar. *Critique of judgment*. Hackett Publishing, 1987.

[12] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE CVPR*, pages 419–426, 2006.

[13] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *IEEE PAMI*, 20(3):226–239, 1998.

[14] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[15] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen. Exploring relationships between audio features and emotion in music. *ESCOM*, 2009.

[16] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213. IEEE, 2011.

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Multimedia*, pages 83–92. ACM, 2010.

[18] M. L. Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pages 22–28. Desire Network, 2010.

[19] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE ICCV*, pages 1784–1791, 2011.

[20] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1):1–23, June 2011.

[21] A. K. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In *IEEE ECCV*, pages 1–14. 2010.

[22] M. D. Mumford. Where have we been, where are we going? taking stock in creativity research. *Creativity Research Journal*, 15(2-3):107–120, 2003.

[23] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE CVPR*, pages 2408–2415, 2012.

[24] A. Newell, J. Shaw, and H. A. Simon. *The processes of creative thinking*. Rand Corporation, 1959.

[25] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *IEEE CVPR*, pages 33–40, 2011.

[26] M. Redi and B. Merialdo. Saliency moments for image categorization. In *ACM ICMR*, 2011.

[27] M. Redi and B. Merialdo. Where is the interestingness? retrieving appealing videoscenes by learning flickr-based graded judgments. In *ACM Multimedia*, pages 1363–1364, 2012.

[28] J. Rigau, M. Feixas, and M. Sbert. Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. *Computational Aesthetics in Graphics, Visualization, and Imaging*, 2007.

[29] F. Sparshott. Basic film aesthetics. *Journal of Aesthetic Education*, 5(2):11–34, 1971.

[30] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology*, 123(4):394, 1994.

[31] R. W. Weisberg. *Creativity: Beyond the myth of genius*. 1993.

[32] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *IJCV*, 96(3):353–383, 2012.