Broadcasting oneself: Visual Discovery of Vlogging Styles

Oya Aran, *Member, IEEE,* Joan-Isaac Biel, and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—We present a data-driven approach to discover different styles that people use to present themselves in online video blogging (vlogging). By vlogging style, we denote the combination of conscious and unconscious choices that the vlogger made during the production of the vlog, affecting the video quality, appearance, and structure. A compact set of vlogging styles is discovered using clustering methods based on a fast and robust spatio-temporal descriptor to characterize the visual activity in a vlog. On 2268 YouTube vlogs, our results show that the vlogging styles are differentiated with respect to the vloggers' level of editing and conversational activity in the video. Furthermore, we show that these automatically discovered styles relate to vloggers with different personality trait impressions and to vlogs that receive different levels of social attention.**

I. Introduction

AMONG the vast and diverse collection of videos in YouTube, much of the content is amateur, user-generated videos [15]. With recent developments in video production tools (e.g. phones with cameras, webcams, and editing software), it is now easy to create and post videos to these sites. This fact engages more people to use this technology as a new way of online communication with remote audiences. Based on this technology, as a natural extension of text-based blogging, video blogging (vlogging) has emerged [13]. Video bloggers (vloggers) record themselves and share their vlogs on social media sites, and in comparison to text-blogging, vlogging provides a richer environment with the use of the video medium. The richness of expression and the diverse content of vlogs, spanning issues from personal diaries to commentaries on everyday life or world events, makes them appealing to a wide audience.

One can find many different types of vlogs, as people post on different subjects using a variety of video creation techniques. One of these types, conversational vlogs, is the focus of this study. Conversational vlogs present a monologue-like setting in which vloggers display themselves in front of the camera and talk. Although conversational vlogs are asynchronous and recorded as monologues, they establish conversations between the vloggers and their audience. In a this setting, apart from what is being said (the verbal channel), the nonverbal channel becomes equally important [23].

Previous work in the study of conversational vlogs has identified the correlates between certain individual behavioral cues of vloggers and the impressions that people make about them, and have also addressed the prediction of vlogger impressions using supervised methods [5], [4]. In this article, we are interested in automatically analyzing the visual content of conversational vlogs to identify different *styles* people use to present themselves and their ideas, and how these different styles are perceived by their audience. By discovering these styles, which we call vlogging styles, we aim to identify common communication elements used in vlogging. From the social computing perspective, the study of vlogging styles is important for the understanding of vlogger behavior, because in contrast to the works above, puts the focus on the overall composition of the vlogs that result from the interplay of vloggers' choices and behaviors. From the multimedia perspective, styles can be used for vlogger characterization, and can enable functions like vlog indexing based on styles, and vlog collection browsing and recommendation based on content similarity. In this context, the use of unsupervised methods is motivated by the discovery task, compared to predefined classification tasks.

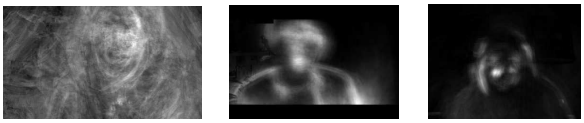
A vlog, used as a way of communicating with a wide audience, is the end result of a video creation process. While the main message of a vlog is communicated by the verbal content and the nonverbal behavior of the vlogger, other conscious and unconscious choices that the vlogger makes during this process convey side messages. Some of these choices affect the video quality: the vlogger selects a webcam, which sets the resolution and the frame rate; a physical place, which affects the lighting. Some other choices affect the video appearance: vloggers decide on the framing (whether their face, upper body, whole body, or something else will be in the camera focus); the places they chose to record the videos set the background (e.g. outside or inside, a tidy or a messy room, etc.); they choose to use a moving or a stationary camera. Finally, the recorded video is either published as is or edited to combine other shots such as images, other video segments, introductory and closing sequences, credits, etc. [6], [24], [28]. All these choices are reflected in the vlog and conveyed to the audience as a communicative signal. *We define the combination of all these choices, which results in the final vlog people share, as a vlogging style.*

To illustrate the concept of vlogging style, we discuss three different vlogs with different visual and temporal character-

This work has been supported by the Swiss National Science Foundation (SNSF) Ambizione fellowship under the project "Multimodal Computational Modeling of Nonverbal Social Behavior in Face to Face Interaction"(SOBE) and the SNSF National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

O. Aran and J.I. Biel are with Idiap Research Institute, Martigny, Switzerland (email: oya.aran@idiap.ch; jibiel@idiap.ch).

D. Gatica-Perez is with Idiap Research Institute, Martigny, Switzerland and École Polytechnique Fédérale de Lausanne, Switzerland (email: gatica@idiap.ch)



(a) (b) (c)

Fig. 1. Example wMEIs from sample vlogs.

istics and in Figure 1, for each of these vlogs, we show the spatio-temporal descriptors, the weighted Motion Energy Images (wMEIs), that we use in this study. The details of this descriptor is presented in Section III-A. Figure 1(a) shows the wMEI of a highly edited vlog in which the vlogger is very active. The resulting wMEI is very cluttered, showing motion in a majority of the pixels in the frame. The corresponding vlog in Figure 1(b) is less edited and the vloggers activity is moderate. The resulting wMEI shows the contours of the head and upper body of the vlogger. For the vlog represented in Figure 1(c), the vlogger is stationary and the vlog contains no editing. As a result, the wMEI includes few pixels with motion, mainly in the face area.

In this study, we focus on the analysis of vlogs from the perspective of visual vlogging styles. We treat each visual aspect of a vlog, i.e., physical elements affecting visual appearance, including resolution, lighting, framing, motion, background, etc., as a signal and we capture the combination of these signals based on the dynamic visual information extracted from the vlog. Our aim is to discover different styles based on clustering methods and to analyze how the found clusters correlate to physical characteristics of the vlog, social characteristics of the vloggers themselves, and indirect indicators of how the vlogs received by their audience. Specifically, we make four contributions:

- We analyze a 120-hour collection of vlogs from YouTube to discover the underlying structure using unsupervised learning techniques. We show that the clusters found by our approach correspond to different vlogging styles, and that the variety of vlog content can be explained by a relatively small number of visual prototypes that reflect many choices on the vloggers' part.
- We collect new crowdsourced annotations on the physical video properties of vlogs, for investigating the links between these elements and the discovered vlogging styles. The annotations include questions on the image resolution, level of motion, amount of framing, lighting, background, and place of recording. These annotations also enable us to identify common properties of vlogs in terms of physical video properties.
- We propose a holistic approach, by analyzing the vlog as a whole, with both conversational and non-conversational aspects, considering that the vlogging style manifests itself not only in conversational context, but also in non-conversational parts. Following this approach, we show that the physical aspects of the vlog production, such as the number of conversational shots, level of editing, resolution, and amount of motion are represented differently in different vlogging styles. We show that the discovered vlogging styles correspond to vloggers with specific personality trait impressions. We also show that the vlogging style significantly relates to the number of views of the vlogs. This finding is interesting because it suggests that the prototypes (styles) extracted with our framework reflect personal traits and viewer's responses at the population level.
- We use a spatio-temporal representation of videos, weighted Motion Energy Images (wMEIs), as descriptors of a vlog. wMEIs have been recently proposed as descriptors of conversational-only videos [5], [30]. In this work, we show that wMEIs can be used as descriptors for edited videos with both conversational and non-conversational parts. As a fast and robust feature extraction method, wMEIs are suitable to use on large scale data. Moreover, a wMEI provides a single image summary of the vlog's visual content, which is not only suitable for automatic analysis, but also meaningful to human observers.

The paper is organized as follows: In Section II, we discuss the most closely related work. We present the details of the wMEI representation, feature extraction and clustering methods that we use for our analysis in Section III. In Section IV, we present the vlog data and annotations. The experiments and results are given in Section V. We conclude and discuss possible future directions in Section VI.

II. Related Work

We discuss related work in two domains: video classification and mining, and vlog analysis.

A. Video Classification and Mining

Video classification and mining is a large domain, and several extensive surveys exist. In [9], a survey presents text, audio, and video based approaches for video classification. In [29], spatio-temporal video retrieval is reviewed. In [33], a review on concept based video retrieval is presented. A more recent survey on video indexing and retrieval is given in [20].

In video classification, the task is to classify a given video into one of the categories. The categories can be broadly defined, such as movie or news, or can be more specific, e.g. identifying various types of sports videos. While initial works used professional video databases or home videos, with the availability of web video resources, recent works explore this content for classification and retrieval [39], [38]. In comparison to professional videos, web videos are rather diverse in terms of format, quality, and subject, which makes the classification and retrieval tasks challenging. Among the many categories defined for web videos, vlogs represent a specific category, in which people post videos to communicate their thoughts and experiences. Although vlogs have been considered as one of the categories in some video classification tasks [39], they have not been explored in detail as a specific category.

While most works on video classification and mining perform their analysis based on the visual content, other works use text, audio, and metadata features as well, either alone or in combination with the visual ones. In most of the approaches, video sequences are treated as collections of still images,

and the visual analysis is based on finding keyframes and extracting low level features from them [29]. This is based on the assumption that the frames in a shot, i.e. the collection of frames within a single camera action, has strong correlations with respect to the full content, therefore only few keyframes can be used to represent the whole shot. These approaches ignore the spatio-temporal characteristics of videos. Although several works use visual spatio-temporal information, the considered tasks are more specific than video categorization (e.g. action recognition [27], [36]). The wMEIs used in our work represent spatio-temporal characteristics of the videos in a 2D grayscale image. wMEIs also provide a computationally efficient and robust way for web video feature extraction.

The features that can be extracted from the visual content can be categorized as shot/keyframe based, object based, or motion based [20]. Motion features, which are our primary focus, have been found to better represent certain semantic concepts in comparison to other feature types. Optical flow and frame differencing are among the used techniques to estimate video motion. In one early example [21], it was shown that for the task of classifying a video as news or sports, optical flow and frame differencing provide similar results, with frame differencing being computationally more efficient. We make use of this result in our work and use frame differencing for detecting moving pixels in each frame, as the first step of wMEI computation. Most works in the literature use shot-based analysis. Instead, we use a representation of the complete vlog via wMEIs, without relying on shots. Our assumption follows the fact that, despite the huge variety of vlogs, conversational vlogs have a structure and the dominating structure will be reflected in the wMEI.

Several works discuss the effects of style of filmmakers on movies and computational approaches to detect them [34], [3], [2]. In addition to movies, a framework for news video indexing based on style analysis is presented in [32]. In their approach, style is grouped into four main features: layout, capture, content, and context. Layout contains the style elements such as shots, transition edits, special effects; capture contains the elements related to the sensors used such as distance, angle, motion; content includes the people, objects, and settings in the video; and context includes the semantic concepts in the video, such as indoors, outdoors, commercial. While the above works consider style in professionally produced videos, in this study we focus on amateur videos. This difference is important as in professional videos, the style elements are mainly conscious choices of the producers, whereas in amateur videos, both conscious and unconscious choices determine the style. Moreover, elements such as low resolution, low quality, bad lighting are often observed in amateur videos, making automatic processing more challenging. An example of early work on analysis of amateur video is [14], which showed trends of non-professional films, but did not address the specific vlog genre.

B. Analysis of vlogs

The establishment of vlogging as a popular genre of online video has generated interest in new media interested on understanding this type of social media. Some research has analyzed the use of vlogging as a means to develop and maintain social relationships on the basis of the video sharing [26]. Other works have studied the process of self-presentation [18] and the experience of self-awareness [37] generated by creating and viewing vlogs, and have also studied how vloggers react to hostile comments [25]. As a main common limitation, these works relied on manual inspection of vlogs and so were reduced to the analysis of small samples of videos.

The current paper contributes to initial works on automatic processing of vlogs with techniques that scale to large amounts of data. In [6], Biel and Gatica-Perez studied vlogs from the perspective of behavioral analysis, proposing a scheme for identifying conversational shots from vlogs, and useing automatic techniques to extract nonverbal cues of audio and video to characterize vloggers' behavior. This investigation provided initial evidence that the extracted nonverbal cues are significantly correlated with the mean level of attention that videos receive (measured by the number of video views), which can be seen as a proxy of the type of impressions that audiences build from watching videos. In [5] and [4], we studied personality impressions from vloggers and investigated how they are associated to a set of automatically extracted nonverbal cues. The results were consistent with [6] in that nonverbal cues associated to high levels of attention were also associated to personality traits that are often socially seen as more desirable. In addition, these works showed that the nonverbal cue representation of vloggers is useful to automatically predict the personality judgments with promising performance. More recently, an investigation incorporating judgments of attractiveness and mood, to obtain a richer characterization of vloggers was presented in [7].

In this study, we assume that all the conversational and non-conversational elements contribute to the vlogging style, and analyze the vlogs as a whole. While our previous works focused on the analysis of the conversational aspect of vlogs [5], [6], [7], [4], vlogs contain also non-conversational content which include still images, text, short video segments, among others. A manual analysis of a sample of vlogs showed that 45% of conversational vlogs contained some type of non-conversational video snippet, which in most cases appeared in the middle of the video [6]. Although the use of openings, middle shots, and endings in conversational vlogs seem to be less frequent than in other types of online video as analyzed in [24], we consider that the non-conversational content is necessary to understand the vlog as a result of the vloggers' creative expression and style.

Our current study extends and contributes to the existing work on vlogs by considering the non-conversational aspects of a vlog as well, in addition to the conversational ones. We propose the use of a descriptor suitable to characterize both conversational and non-conversational vlog content. Through the use of these descriptors in a clustering framework, we discover vlogging styles, which are characterized by different visual aspects of a vlog, such as the amount of motion, the level of editing, and so on.

III. Discovery of Vlogging Styles: Our Approach

We follow an unsupervised approach to discover different vlogging styles in a YouTube vlog collection. From a dataset of vlogs (see Section IV), we calculated the weighted motion energy images as descriptors of the visual activity in the vlog and extracted features via principal component analysis. We then applied k-means clustering to group similar vlogs. Sections III-A, III-B and III-C explain the details of the activity description, feature extraction, and clustering, respectively.

A. Activity Description

For representing and analyzing human action, Bobick and Davis [8] proposed the Motion Energy Image (MEI) and the Motion History Image (MHI), as ways to summarize the spatio-temporal content in a single image. MEI is a binary image showing the pixel-wise location of the motion, in contrast MHI is a grayscale image showing both the location and the direction of the motion. Both MEI and MHI were proposed as motion templates to describe short motion, and have been widely used for human action recognition [8], [35]. As an extension to MEI, in [19], the Gait Energy Image (GEI) was proposed as an alternative, noise-robust, and computationally efficient representation of human action. A GEI is calculated by accumulating the aligned binary human silhouettes in each frame of an action. Normalized by the total number of accumulated frames, the image presents a compact grayscale representation of human action. These approaches have been used in human action analysis, such as person recognition [19] and gait recognition [40].

In this study, we use a modified version of the motion energy image, called "Weighted Motion Energy Image" (wMEI) [5], [30]. A wMEI represents the dominant motion regions in a video, and is suitable to be used as a visual template for long duration videos. It is a grayscale image describing the location along with the intensity of motion throughout the video.

A wMEI contains the accumulated motion information and is calculated as

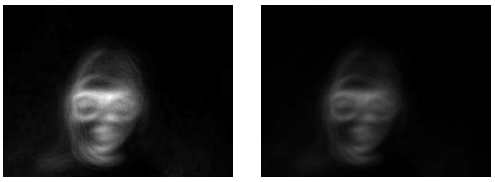
$$wMEI_i(x,y) = \frac{1}{N_i}\sum_{t=1}^{T_i}(D_i^t(x,y)), \tag{1}$$

where (x,y) denotes a pixel, D_i^t is a binary image that shows the moving regions for video i at frame t, N_i is the normalization factor, and T_i is the total number of frames. To obtain the actual grayscale image, the resulting wMEI should be mapped to grayscale intensity levels, as Equation 1 produces values between zero and one.

To obtain the binary image D_i^t, we used frame differencing and thresholding. For this process, the color frames are first converted to grayscale. After applying frame differencing, the moving pixels are identified using a fixed threshold:

$$D_i(x,y) \quad = \quad \begin{cases} 1, \ if \ |V_i^t(x,y) - V_i^{t-1}(x,y)| > \Phi \\ 0, \ otherwise \end{cases} \tag{2}$$

where (x,y) denotes a pixel, V_i^t denotes the grayscale video frame at frame t, and Φ is the fixed threshold that is used to detect moving pixels. D_i^t corresponds to the binary motion information for video i at frame t, as used in Equation 1, which is accumulated to form the wMEI for that video.



(a) (b)

Fig. 2. Example wMEIs with different normalization techniques. Normalization with respect to (a) the maximum accumulated pixel value, and (b) the duration of the video

The selection of the fixed threshold depends on several factors. The threshold should be high enough to eliminate most of the spurious motion effects resulting from low video quality, compression, etc., and low enough to allow the identification of actual moving pixels. For this study, we empirically determined the value of this threshold as 30 on a set of randomly selected videos, which we found to be suitable for the generality of the vlogs

Unlike motion energy images, a wMEI is not a binary image. A wMEI describes the motion throughout a video as a grayscale image, where each pixel's intensity indicates the amount of visual activity in that pixel. Brighter pixels correspond to regions where there is more motion. wMEI is a general purpose representation of the video content and does not require any alignment or binary silhouette extraction. Moreover, as for the normalization factor, a wMEI can be normalized with different approaches to emphasize different aspects of motion. For example, it can be normalized with respect to the maximum accumulated pixel value ($N_i = max_{(x,y)}(\sum_{t=0}^{T}(D_i^t(x,y)))$) or with respect to the video duration ($N_i = T$). In the former approach, the resulting image is not affected by the video duration or by the stationary parts of the video. For the same video, the second approach will generally produce a darker wMEI than the wMEI of the first approach. This is valid as long as the maximum accumulated pixel value is smaller than the video duration. If there is at least one pixel in the wMEI which accumulates motion in every frame of the video, then the normalization factor will be equal in both approaches, producing the same wMEIs. Figure 2 shows the two wMEIs of the same vlog normalized with two different normalization factors. In the first case, the resulting wMEI has higher dynamical range and is not affected by the video duration or by the stationary parts of the video. To illustrate this, assume that we have two vlogs where the second one is the same as the first one with the addition of a still image that is shown for some time at the end of the video. As there is no motion in the added part with the still image (except for the first frame), the resulting wMEIs based on normalization with maximum accumulated pixel value would make almost no difference between these two vlogs. However, if the duration of the video is used for normalization, the wMEI of the second video will be significantly darker. For the rest of the discussion in this paper, we will use the first normalization technique in order to emphasize the non-stationary parts of the vlogs.

In Figure 3, we show wMEI examples corresponding to

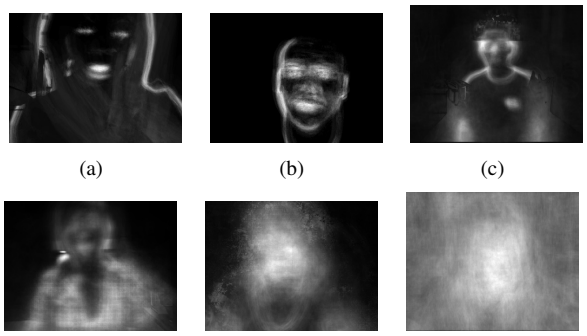


Fig. 3. Sample wMEIs of conversational vlogs: (a)-(b) limited vlogger movement, (c) arm movement, (d)-(f) significant movements

six conversational vlogs from YouTube. The duration of these vlogs are between 38 seconds to seven minutes. Despite the differences, these wMEI examples reflect one common setting of this video genre: a single person speaking in front of a camera. In these grayscale images, one can notice that the wMEIs contain the silhouette of a person's upper body or face, with different characteristics. The silhouette is very clear when the person's movements are limited (see Fig. 3(a) and Fig. 3(b)). Different framing styles, i.e. closer to or farther from the camera, centered or not, are also apparent. In Fig. 3(c), one can see an example when a vlogger clearly moves his arms. In all these cases, the wMEI captures the movements in the contours of the body and face. The more the person moves throughout the vlog, the blurrier the silhouette (Fig 3(d) - 3(f)).

As the wMEI is calculated using the motion information, it is hardly affected by stationary frames of the video. While edited shots with activity is reflected in the resulting wMEI, edited shots with still images have almost no effect. The motion is detected using frame differencing, therefore it does not directly differentiate between the camera motion and the vloggers activity. However, while the camera motion causes motion in a majority of the pixels in the camera view, the vlogger's activity only affect the pixels that correspond to the vlogger. The clothing of vlogger also has an effect in the resulting wMEI: motion can not be detected on smooth textures. For example, in Figure 3(d), although the vlogger is active and most of this activity is reflected in the wMEI, no motion is detected on the skin-colored neck region. Similar effects can be seen with single colored clothing.

The wMEI exploits the general structure in a conversational vlog, in which the vlogger is in front of the camera and talks in a monologue like fashion. By treating every visual aspect of a vlog as a communication signal, a wMEI captures these signals based on the motion information. The resulting grayscale image is a representative of the degree and style of conversation in the vlog. If the video is hardly edited and the vlogger is mainly stationary with limited movements (i.e. sitting), the wMEI would be a uniform image, only containing the contours of the vlogger's body or the face, resulting from the movements during speaking (i.e. facial movements, body leaning). On the contrary, for an edited video, with a very active vlogger, one would expect a very cluttered wMEI. With these properties, from the perspective of vlogging style, which we define based on the activity of the vlog (both in terms of the vlogger behavior, and the level of video editing and varying shots), wMEI is a suitable representation choice.

We have used the wMEIs as visual descriptors of video content in our previous studies as well. In [30], to analyze emergent leadership behavior from nonverbal cues, we use entropy, mean, and median, extracted from wMEIs of participants in a meeting. In [5], we computed wMEIs for conversational vlog videos and extracted features such as entropy, mean and median to analyze the personality impressions of the vloggers. The experiments showed that among all features, including several audio and visual features, the wMEI based features have the highest correlations with several personality trait impressions. These works show that even simple features extracted from wMEIs provide valuable information on the visual video content for several other tasks. In the current study, we extend the use of wMEIs for vlogs containing both conversational and non-conversational parts, and show that it a robust descriptor for analyzing vlogs in a broader range.

B. Feature Extraction

Each vlog in our dataset is first processed to obtain the wMEIs (see Section III-A). As YouTube users upload videos with different resolution (see Section IV-A), the resulting wMEIs are also of different sizes. To have a fixed image size, we resized each wMEI to 320×240 (width×height) pixels, which is the median resolution of the videos in our dataset.

From the normalized and resized wMEIs, we extracted features based on Principal Componenet Analysis (PCA). Prior to PCA, the wMEIs are first centered by subtracting the mean image of the whole dataset from each wMEI. The eigenvectors corresponding to the eigenvalues explaining 95% of the variance are used to map the original wMEIs, which results in a new feature space with 90 dimensions.

Methods other than PCA can also be used to extract features from the wMEIs. Discrete Cosine Transform (DCT) and Histograms of Oriented Gradients (HOG) are two other techniques that extract features describing both the spatial structure and the appearance of the image. Alternatively statistics on the wMEI pixel intensities can be used. These include simple image statistics such as entropy, mean, median or a histogram analysis on the wMEI, representing the distribution of the pixel intensities. However, these representations only contain the appearance information, discarding the spatial structure. For comparison purposes, we performed clustering with other features as well, including Discrete Cosine Transform (DCT), Histograms of Oriented Gradients (HOG), several IMage Statistics (IMS), and image HISTogram analysis (HIST). The details for each technique are summarized below:

- DCT: We calculated the 2D DCT of the wMEI and collected the coefficients corresponding to the top left half of the first 50×50 block (including DC) in zigzag pattern from the top left corner, resulting in a feature vector of size 1275.
- HOG: We calculated the HOG features as described in [11], using the following parameters: 16×16 cell size, 2×2 block size, and nine bins in each histogram. This

results in a feature vector of size 9576. We reduce the dimensionality of this feature vector to 860 by applying PCA with 95% of the variance explained.
- IMS: We used the entropy, mean and median values extracted from each wMEI as features, forming a feature vector of size three.
- HIST: We calculated the image histogram of the wMEI using grayscale intensities as bins and obtained a feature vector of size 256.

C. Clustering

We use K-means algorithm for clustering the feature vectors extracted from wMEIs. K-means is a well known clustering algorithm [22] and extensively used in diverse fields, including video mining, video shot clustering, and key frame clustering [9], [29]. Moreover, the close connection between PCA and K-means makes it a suitable choice as a clustering algorithm for this task. It has been shown that the principal components found by PCA provide a continuous solution to the discrete cluster indicators in K-means clustering [12]. It has also been shown that reducing the dimensionality of the original data using PCA maps the data to the most discriminative subspace. Thus, applying K-means on the PCA reduced subspace is more effective than applying K-means on the original space.

K-means algorithm starts with k initial cluster centers and iteratively assigns each observation to the nearest cluster center, updating the cluster centers at each iteration. Thus, the clustering result is subject to change if a different set of initial cluster centers is used. For our dataset, the clustering result is found to be robust to the choice of initial cluster centers: for 10 different k-means runs with $k = 3$, initialized with random cluster centers, we obtained exactly the same clustering results.

IV. Data and Annotations

A. Data

We use a dataset of conversational vlogs downloaded from YouTube, originally presented in [6]. The data was obtained by querying videos from YouTube using three possible keywords: vlog, vlogging, and vlogger, and then by manually selecting the conversational vlogs among them. The dataset contains 2268 single-person videos from 469 users with metadata (title, description, duration, keywords, video category, date of upload, number of views, and comments), corresponding to a total of more than 160 hours of video. There are one to eight vlogs per user. The median duration of vlogs is 3.4 min with a median frame rate of 30 fps. The resolution of the videos (width×height) varies with 36 variants, in the range of 640×480 and 160×120, with a median of 320×240.

B. Annotations

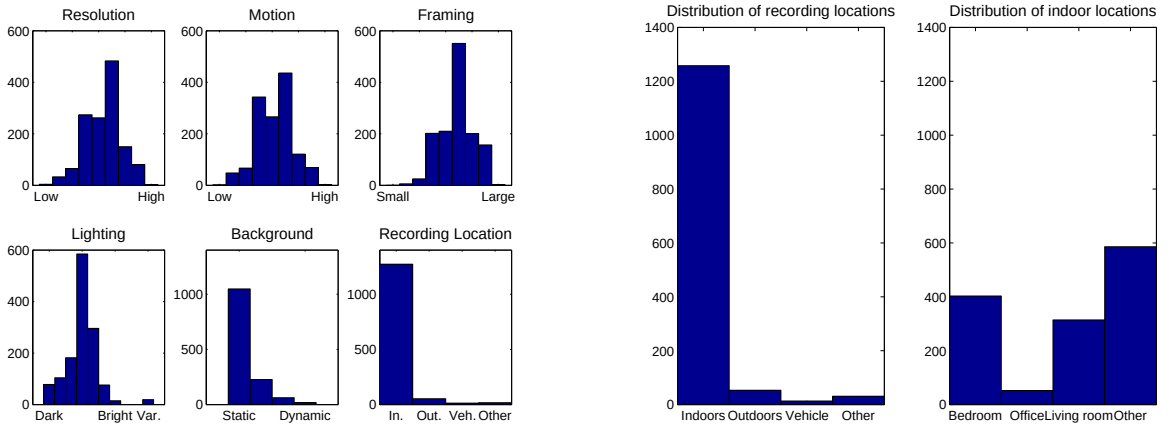
We used several annotations performed on this dataset. These annotations, obtained either manually or automatically, were meant to study the vlogs along different dimensions, and used in the analysis and validation of our framework.

There are two automatically obtained annotations for each vlog: first is the identification of conversational and non conversational shots; second is the extraction of a one-minute conversational segment as a representative of the vlog. Moreover, there are two types of manual annotations collected from external reviewers via crowdsourcing: the first type includes several aspects of vlogs, such as resolution, amount of motion, and framing; the second one is the personality impressions about the vloggers. In addition, we used the number of views from the video metadata as a measure of social attention. The details are described below.

1) Vlog shot analysis: To objectively assess the vlog content in terms of editing and level of conversation, we use results of the analysis presented in [6], which automatically processed vlogs, detected shots in the video, and found conversational and non-conversational shots. The details can be found in [6]. This analysis provides the number and duration of the conversational and non-conversational shots in each vlog. It shows that 55% of the vlogs contain a single shot, that the mean number of shots is 3.9, and the conversational parts correspond to 89% of the vlog duration on average.

2) Extracting a one-minute conversational representative: In [5], the vlogs in the dataset were processed to automatically extract the first one-minute conversational segment. Although the vlogs in our dataset are manually selected as conversational, we noticed that in some vlogs the conversational parts are very short. Nevertheless, we were able to extract a one-minute conversational segment for a majority of videos (2182 vlogs out of 2268).

3) Vlog physical video properties: We designed and collected a new set of annotations for several aspects of the vlog using Amazon's Mechanical Turk (MTurk) crowdsourcing platform [1]. The annotations were performed on more than half of the whole dataset (1355 vlogs). The subset contains the 442 vlogs that are annotated for personality (see Section IV-B4) and the remaining vlogs are selected semi-randomly, ensuring that each user is represented proportionally (the number of vlogs from each user in the subset is proportional to the whole dataset). The annotators were asked to watch the first minute of a vlog and answered several questions related to physical aspects of the vlog. We asked about the quality of the video resolution, the level of motion, and the amount of framing occupied by the person in the video, obtaining answers in a 5-point Likert scale (1-low to 5-high for resolution and motion, 1-small to 5-large for framing). We also asked a question about the lighting in the video, with possible answers being constant - dark, constant - normal, constant - bright, and variable. Another binary question asked for the background, whether it is static or dynamic. We also asked about the place of recording: indoors (bedroom/living room/office/other), outdoors, vehicle, other. Each vlog was annotated by three annotators. We applied no restrictions regarding the country of the annotator. The consensus is obtained by calculating the average for the questions with ordinal answers (resolution, motion, framing, background) and majority voting for the questions with discrete answers (place). For lighting, we used majority voting to decide whether it is constant or variable, and calculated the average for the cases with constant lighting. The histograms of the annotations are shown in Figure 4. Based on these histograms, one immediate



(a) (b)

Fig. 4. (a) The histograms of vlog video annotations on the quality of resolution, level of motion, amount of framing, background, lighting, and place of recording. (b) The details of the annotations on the place of recording.

observation is that, a majority of the vlogs is recorded indoors, mostly with a static background and constant lighting.

4) Personality impressions: We use manual annotations for personality, which were presented in [5]. The annotations were performed on a subset of the data and the annotated subset was selected such that there is one video per vlogger, and there is at least a one-minute conversational part in the video. The subset contains 442 vlogs of which 47% (208) corresponded to male vloggers and 53% (234) to female vloggers. MTurk was also used in order to obtain zero-acquaintance judgments of personality. The annotators answered the Ten-Item Personality Inventory (TIPI) questionnaire, a 10-item measure of the Big Five personality dimensions [17], about the person appearing in the vlog. Each vlog was annotated by five annotators. More details about the annotations can be found in [5]. We use these annotations to analyze how the found clusters relate to the personality impressions of the vloggers.

5) Social attention: In vlogs, the social attention measured by the number of views is the aggregate result of people watching vlogs, and is a proxy that accounts for the multiple impressions that are built from a single vlogger. We use the metadata of the vlogs and use the log number of views to indicate the social attention for each vlog, as presented in [6].

V. Experiments and Results

The experiments presented in this section aim to show that the clustering approach that we present is able to capture the structure and characteristics of vlogs to discover vlogging styles. We perform our experiments on the dataset presented in Section IV-A and use manual and automatic annotations, presented in Section IV-B, to show the validity of clustering and to analyze further relations of vlogging styles with physical video elements and other aspects such as personality impressions and social attention.

In the next section, we present different clustering results with different feature vectors and number of clusters. Then, based on the clustering using feature vectors obtained by PCA, we analyze the clustering results with respect to the annotations. In the light of these results, we discuss whether the found clusters relate to different vlogging styles. Finally, we investigate the relation between vlogging styles, personality impressions of vloggers, and social attention.

A. Clustering vlogs: discovery of vlogging styles

To analyze the effect of feature extraction and the number of clusters, we performed experiments using k-means clustering with different feature vectors and also with different number of clusters. Figure 5 shows the mean images of each cluster for k=3 for each of the feature sets. To facilitate comparison, we set the cluster IDs such that the smaller the ID, the higher the mean intensity of the mean image of the corresponding cluster. As can be seen from these images, the mean images of each cluster resemble the silhouette of a person, with different degrees of clarity.

A general look at Figure 5 indicates that three different levels of conversational activity can be extracted by all methods. Furthermore, PCA and DCT produce very similar clustering results as judged by the mean images and the distribution of data samples over clusters. HOG also produces similar clustering results, although with a more balanced distribution of data over clusters. It is important to note that PCA and DCT calculations are holistic, i.e. computed on the whole image, whereas HOG calculations are block based. The clustering results using IMS and HIST features are different than other feature sets, both in terms of the mean images and also the cluster distribution. With both IMS and HIST features, k-means clustering produces a very big first cluster and a relatively small third cluster, which includes very dark wMEIs, indicating very small motion. On the contrary, PCA and DCT produce a small first cluster, which includes very bright wMEIs (indicating high and distributed motion), and a big third cluster. Both PCA, DCT and HOG features use the spatial information in the wMEIs where as IMS ans HIST features only use statistics on the intensity, without any spatial information. Using IMS or HIST results in extremely unbalanced clusters whereas clusters are more balanced in PCA, DCT

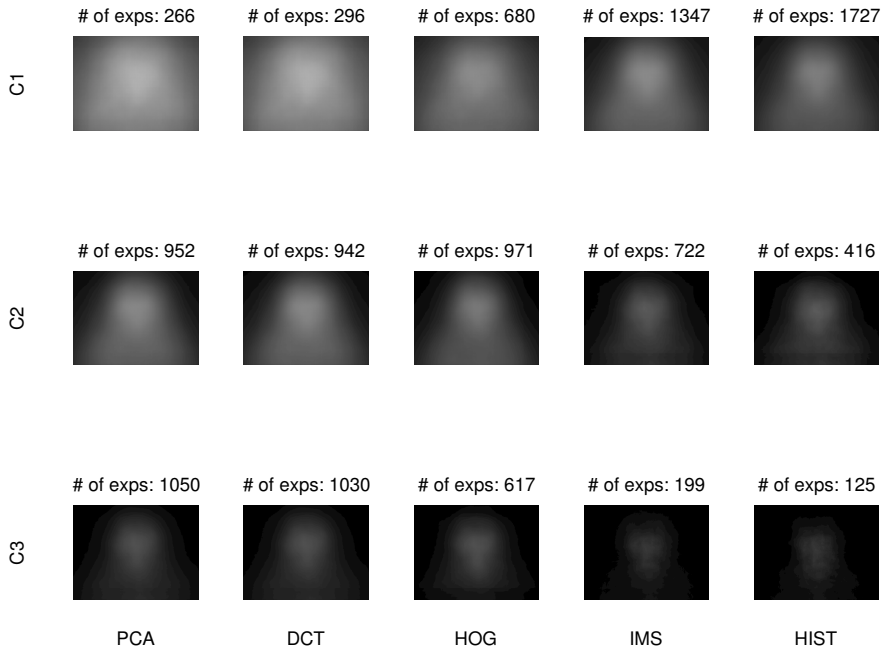


Fig. 5. K-means vlog clustering with k=3 clusters. Mean images of clusters for four feature sets (PCA, DCT, HOG, IMS, and HIST) are shown column-wise. The number of samples assigned to each cluster is shown for each case.

and HOG, with HOG producing the most balanced clusters. While having balanced clusters is not a requirement, having extremely unbalanced clusters indicates that the features do not represent the vlogs well. For the rest of the discussion, we will present the detailed results on PCA features.

As k-means algorithm does not automatically set the value of k, the number of clusters, we need to select the k. First, we experimented with various number of clusters. Figure 6 shows the mean images of each cluster for two, three, four, and five clusters. In all cases, the clusters correspond to a group of videos with different amounts of motion as represented by the wMEIs. As the number of clusters increases, the clusters become more specific and represent a range of vlogs from high to low activity. In order to select a number of clusters, we use the Bayesian Information Criterion (BIC). The BIC scores on 10 clustering attempts for different number of clusters, from 2 to 10, are shown in Figure 7. This experiment shows that 3 or 4 clusters gives the lowest BIC score.

For the rest of the paper, we will base our discussion on the three-cluster case (see Figure 6(b) for the mean images with k=3), as it facilitates the discussion on vlogging styles. The results with four clusters are discussed in the text; detailed figures and tables are omitted here for space reasons but can be seen in the supplementary material.

Figure 8 shows the sample wMEIs in each cluster for k=3. We show the two closest, the farthest, and the middle examples to the cluster centroid. The first cluster contains the brightest wMEIs, thus representing vlogs that have more activity. Although the brighter pixels in the wMEIs indicate the silhouette of a person's upper body, there is activity also in other parts of the frame. The reasons for this can be various. The vlogger could be very active and move a lot. The background could be dynamic. The video could also be highly edited with non-conversational shots or with parts that contain high motion. Possible other reasons include moving camera, low resolution video and/or bad lighting.

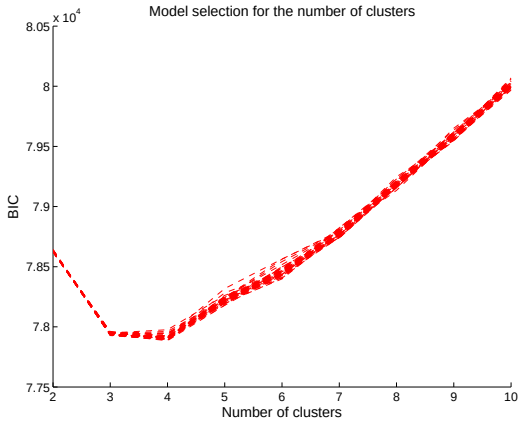


Fig. 7. Model selection for the number of clusters. The averaged BIC score for number of clusters from 2 to 10 is shown. Each line shows a different clustering attempt. The BIC score has a local minima around 3-4 clusters.

Resulting wMEIs of sample vlogs from cluster 1 for different factors affecting the wMEI are represented in Figure 9. Note that the wMEI captures multiple factors that by themselves relate to a specific style to vlog. Some of them relate to the vlogger itself (activity level) and some others to the vloggers choices (e.g. editing) and/or circumstances (e.g. webcam quality). The discovered vlogging styles reflect a combination of all these elements. For example, for a vlog with static background but with significant human activity, the resulting wMEI can be very "blurry". For instance, Figure 9(a) represents a highly edited vlog with a static background. Figure 9(d) represents a vlog without any editing and with a

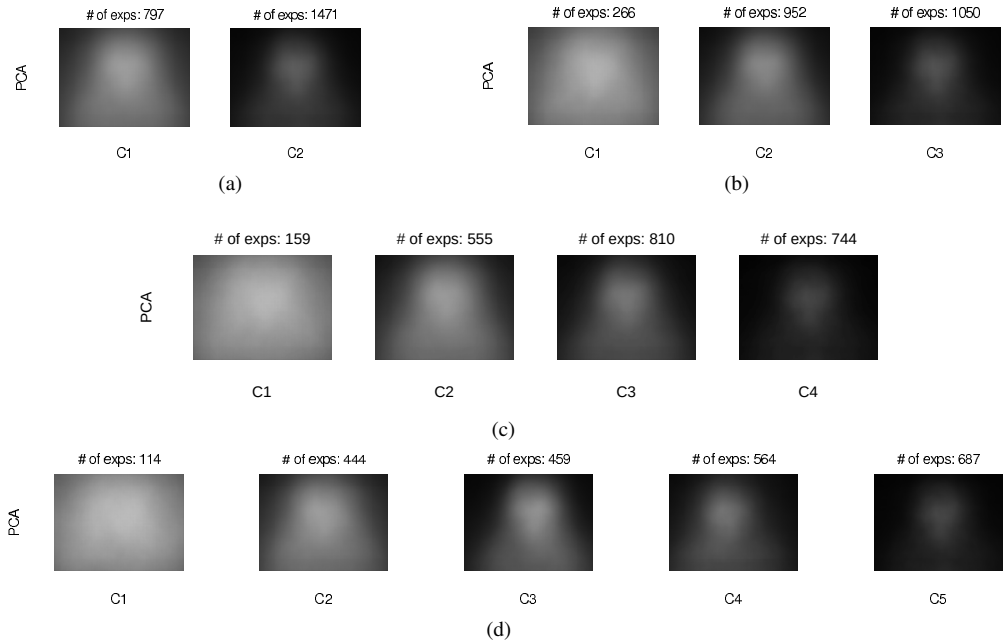


Fig. 6. Mean images of clusters for (a) two, (b) three, (c) four, and (d) five clusters.

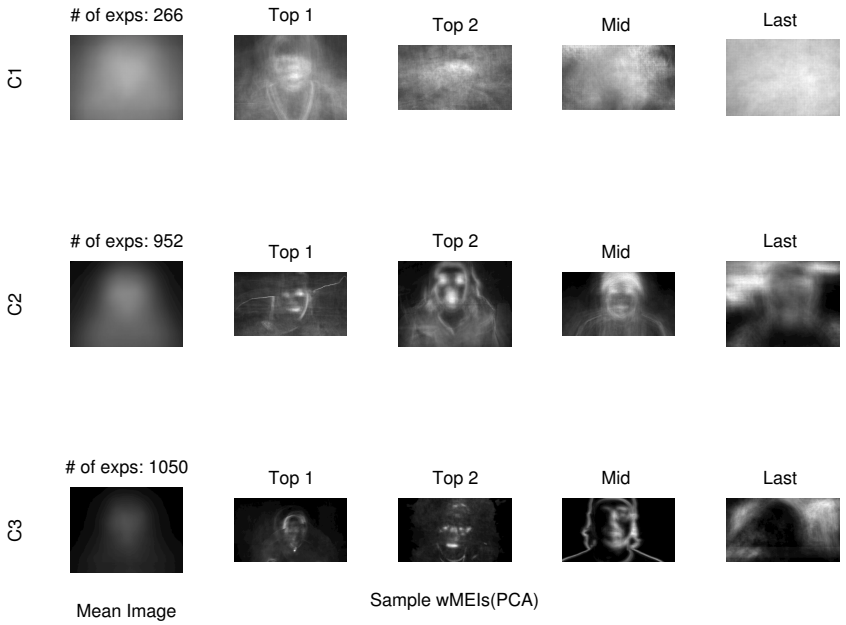


Fig. 8. Clustering results with three clusters using PCA features. Rows correspond to clusters. The first column shows the mean images and the following columns show sample wMEIs for each cluster. Samples shown are the two closest, middle, and farthest samples to the cluster centroid.

static background. Both of these vlogs have a blurry wMEI and are clustered in cluster 1. The second cluster in Figure 8 contains wMEIs that represent the silhouette of a person's upper body, without much motion in the surrounding pixels. The third cluster in Figure 8 contains wMEIs that are darkest and indicate the contours of a person's face or upper body. The vlogs corresponding to this cluster are more likely to be purely conversational vlogs, in which the person is stationary during most of the vlog. Here, stationary refers to a person sitting in front of a camera and talking, without moving too much. The thresholding applied for the detection of the moving pixels eliminates the subtle movements and leaves only the significant ones, mainly reflected in the contours.

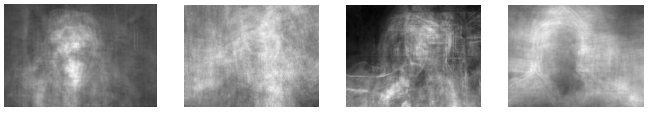


Fig. 9. wMEIs of sample vlogs from cluster 1 for different factors affecting the wMEI appearance. (a) active vlogger; (b) moving camera; (c) editing with non-conversational shots; (d) low resolution/bad lighting.

B. Analysis of vlog clusters

To objectively validate whether the clustering method we apply produces meaningful clusters, we analyze the resulting clusters using the annotations presented in Section IV-B. We

evaluate the clusters with respect to the amount of vlog's conversational content and the number of shots.

For comparison, where necessary, we use two statistical tests. We use a two-sample t-test with a null hypothesis that indicates whether the data in the clusters are independent random samples from normal distributions with equal means and equal or unequal but unknown variances, against the alternative that the means are not equal [31]. We also use a Kolmogorov-Smirnov (K-S) test with a null hypothesis that indicates whether the samples in each cluster are coming from different distributions [31]. The latter does not make any normality assumptions. For each cluster pair, we test the hypothesis that the two clusters are the same (i.e., they have the same mean for the two-sample t-test and, for the K-S test, the samples in the clusters come from the same distribution). For three clusters, three hypotheses are needed to compare all cluster pairs: $H_0 : C_1 = C_2$, $H_0 : C_1 = C_3$, $H_0 : C_2 = C_3$.

As stated earlier, while some vloggers prefer to have conversation-only vlogs recorded in a single shot, others edit their vlogs with conversational and/or non-conversational content. We performed several experiments to evaluate the found clusters with this respect and to show whether these vlogging styles are reflected in the clusters.

The annotations, defined in Section IV-B1, explain the dataset in terms of the conversational and non-conversational shots in the vlogs. Figure 10 shows the statistics of conversational shots for the vlogs of each cluster for three clusters. In cluster 1, 63% of the shots in a vlog are conversational, spanning 77% of the whole video duration. Cluster 3 has a higher ratio both for the number (81%) and the duration (93%) of the conversational shots. These results show that the vlogs in cluster 3 depict more conversational content with respect to the vlogs in cluster 1. Moreover, 63% of the vlogs in cluster 3 contain only conversational shots, whereas it goes down to 48% and 33% for clusters 2 and 1, respectively. Another statistic let us compare the clusters with respect to the number of shots in the vlogs, which we use as a proxy for the level of editing of the vlog. The mean number of shots for vlogs in cluster 1 is 8.3, in comparison to 4.1 and 2.8 for clusters 2 and 3 respectively, which indicates that the vlogs in cluster 1 are clearly more edited. All these figures are statistically significant, according to the two statistical tests, the two sample K-S test and the two sample T-test. In summary, these figures indicate different styles of vlogging: On one hand, the vlogs in cluster 3 correspond to a vlogging style which is more conversational (on average 93% of a vlog is conversational) and less edited (2.9 shots on average). On the other hand, in cluster 1, the vlogs are highly edited (8.3 shots on average) with relatively less conversation (77% on average). We observe the same trend in the four-cluster case as well, where the vlogs in cluster 1 contain significantly less conversational shots than that of clusters 2, 3, and 4.

Another type of annotations, explained in Section IV-B2, extracted a one-minute conversational segment from the vlogs. Comparing the wMEIs of a vlog and its one-minute part can also give insights about the amount of conversational content and editing of the vlog. One could argue that the distance between the wMEI of a vlog and the wMEI of its

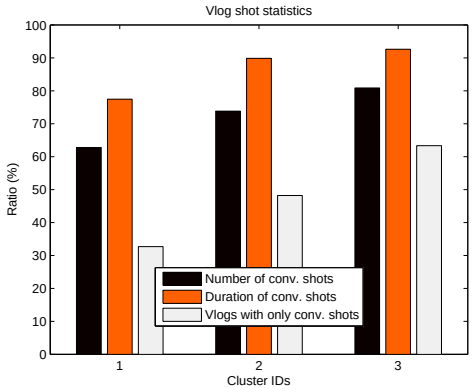


Fig. 10. Conversational shot statistics of the vlog collection in each cluster. The bar graphs shows the mean ratio of the number of conversational shots to the total number of shots and the mean ratio of the duration of conversational shots to the whole video duration. The percentage of the vlogs that contain only conversational shots is also shown for each cluster.

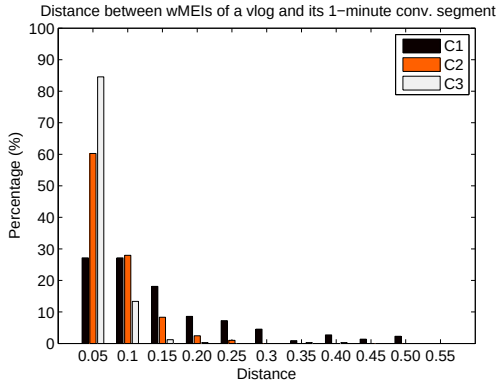


Fig. 11. Distance between wMEIs of a vlog and its one-minute conversational segment.

conversational one-minute counterpart should be small if the video is mainly conversational and not heavily edited, or in other words, if the one-minute conversational part is a good representative of the whole vlog.

To investigate this hypothesis, we compute the distance between the wMEIs of a vlog and its one-minute counterpart by calculating the sum of absolute differences in the normalized wMEIs and normalizing by the total number of pixels. This creates a distance value between 0 and 1 for each vlog. Figure 11 shows the bar graph of the binned distance values in each cluster for three clusters. The x axis indicates the distance in bins and the y axis is the percentage of vlogs that fall into that bin in each cluster, shown as different bars. It can be seen that around 85% of the examples in cluster 3 fall in to the first bin, which has the smallest distance. The mean distance is 0.16, 0.08, and 0.04 respectively for clusters 1, 2, and 3. The four cluster case also indicates that the distances are higher in cluster 1. For both the three and four cluster cases, the two sample t-test and the K-S test rejects the null hypothesis for each cluster pair, with very small P values close to zero.

These results show that, the degree of representativeness of the one-minute conversational part for the whole video is significantly different for the three clusters. For cluster 3, the mean distance between the wMEIs of a vlog and its

one-minute counterpart is the smallest, indicating that the whole video is similar to the one-minute part, hence more conversational. On the contrary, the mean distance for cluster 1 is the highest. For the vlogs in this cluster, the one-minute part does not resemble the whole video. This is the result of either having non-conversational parts or having a very active vlogger who has different communicative styles throughout the video.

Based on the same representativeness assumption, we hypothesize that the two videos, the one-minute segment and the whole vlog, should be assigned to the same cluster if one is a representative of the other. To evaluate this, we processed the one-minute videos to compute the wMEIs and extract the PCA features. Based on the clusters found for the whole vlogs, we assigned the cluster ID of the cluster that has the closest distance between the cluster centroid and the feature vector. The results show that for 91% of the vlogs in cluster 3, their one-minute counterpart is also assigned to cluster 3. The percentages for cluster 1 and 2 is 43% and 60% respectively. These results support our previous results, using a different perspective. The vlogs in cluster 3 are mainly conversational and their one-minute counterpart is a representative of the vlogging style of the whole vlog. Whereas the vlogs in clusters 1 and 2 differ from their conversational counterpart.

C. Video Elements Affecting The Style

In the previous subsection, we have used automatic annotations on shot segmentation and the conversational context of the vlogs to evaluate the clusters and resulting vlogging styles. Those annotations reflect both the conversational aspects and also the level of editing of the vlogs. The analysis in the previous section shows that there is a correspondence between the conversational aspects of each vlog and the discovered clusters, and also between the number of shots in a video and the clusters. The wMEI representation and the clustering based on wMEIs are able to utilize that information to discover vlogging styles. In this section, we further investigate several other video elements that may affect the style. For this purpose, we use the manual annotations obtained through crowdsourcing from Section IV-B3. The annotations contain annotators' judgments on the vlog resolution, level of motion, amount of framing, background, lighting and place of recording.

Figure 12 shows the mean annotation score of each of the properties for the vlogs in each cluster, for k=3. Only the properties with annotations based on ordinal scores, the quality of resolution, level of motion, amount of framing, and background, are shown. The score for the background annotations is mapped to [1,5] (originally scored between [0,1]), for visualization purposes. The statistical test results for the comparison of clusters are given in Table I. First, for the level of motion and the background, cluster 1 is the highest and cluster 3 is the lowest scored cluster. These human judgments of motion are in parallel with the observation that we have obtained based on the wMEIs: the brighter the wMEI, the higher the motion in the video and/or the more dynamic the background, and this can also be seen in the discovered clusters. Second, for the degree of framing, although cluster 3 still has the lowest score, the scores of cluster 2 are slightly higher than those of cluster 1. There is a significant difference only between clusters 1 and 3. Finally, the quality of resolution, on the other hand, shows a different trend, in which the scores are the lowest in cluster 1. Similar trends are observed with the four-cluster case.

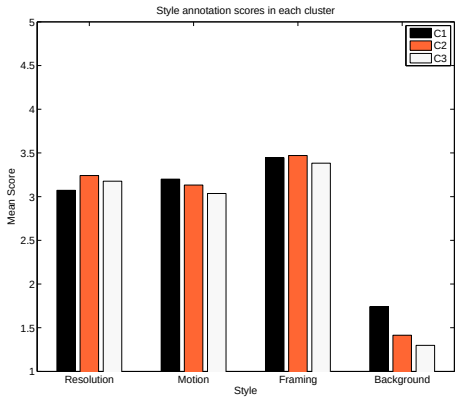


Fig. 12. The annotation scores for vlog video properties in each cluster

TABLE I
STATISTICAL TEST RESULTS ON VIDEO PROPERTIES

	$H_0 : C_1 = C_2$	$H_0 : C_1 = C_3$	$H_0 : C_2 = C_3$
Resolution	$\mathrm{X}^{k\dagger}_{t*}$		
Motion		$\mathrm{X}_{t\dagger}$	$\mathrm{X}^{k\dagger}_{t\dagger}$
Framing		$\mathrm{X}_{t\dagger}$	
Background	X_{t***}	X^{k***}_{t***}	X^{k**}_{t*}

X indicates that the null hypothesis, H_0, is rejected;
k: Two-sample K-S test, t: Two-sample t-test;
***: p<<0.001, **: p<0.005, *:p<0.01, †: p<0.05.

To complement this finding, we also calculated the average frame size (i.e. we used the total number of pixels in each frame to handle videos with different sizes) for the vlogs in each cluster. We see that the results are in parallel with our previous findings: there is a slight increase on the average frame size, from cluster 1 to cluster 3, with cluster 1 being the lowest (statistically significant with the t-test).

These results show that the vloggers in cluster 3 do not move a lot, and use static backgrounds in their vlogs. On the contrary, the vloggers in cluster 1 use more dynamic backgrounds. Moreover, the resolution quality of the vlogs in cluster 1 is lower. For framing, although the average score for cluster 3 is the lowest (i.e. person occupies less space in the frame), it is not statistically significant.

D. Vlogging Style and Personality Impressions

The results in the previous section indicate that the extracted clusters contain vlogs that are different in style. Cluster 3 contains vlogs that are mainly conversational and stationary whereas Cluster 1 depicts higher activity. Finally, the vlogs in Cluster 1 contain less conversational segments and have more editing features and varying shots.

In this section, we investigate whether there is a relationship between these vlogging styles and personality impressions of the vloggers. This hypothesis is motivated, as described

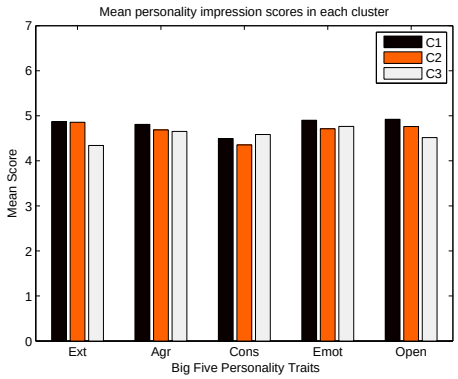


Fig. 13. Mean personality impression scores

TABLE II
MEAN SCORES IN EACH CLUSTER FOR BIG FIVE TRAITS.

	C_1	C_2	C_3
Ext	**4.87**	4.86	4.34
Agr	**4.81**	4.69	4.65
Cons	4.49	4.35	**4.59**
Emot	**4.90**	4.71	4.76
Open	**4.92**	4.76	4.52

in Section II, by works in various social media sources that have found that personality impressions (i.e., personality traits reported by external observers) have correlation with observable features from social media sources [16], [10], [5].

The personality impression scores were obtained on a subset of 442 vlogs (see Section IV-B4 for details). The five external annotations per vlog were averaged, which gives an aggregated personality impression score for the vlog for each of the Big Five traits: Extraversion (Ext), Agreableness (Agr), Conscientiousness (Cons), Emotional Stability (Emot), and Openness to Experience (Open).

Figure 13 shows the bar graphs of the mean personality impression scores for the three-cluster case. In the three-cluster case, the scores for all traits, except for Cons, is the highest for cluster 1 and the lowest for cluster 3 (except for Cons and Emo). For Cons, we observe a reversed bell shape, with cluster 2 being the lowest. We can observe the same trends with four clusters. The actual scores are given in Table II, with bold values showing the highest score for each trait.

We perform two statistical tests, the two-sample t-test and the K-S test, to see whether the personality impression scores of clusters are significantly different. Table III shows the K-S and two-sample t-test results for the three cluster case. For both Ext and Open, the tests reject the hypotheses that clusters 1 and 3, and clusters 2 and 3 come from the same distribution. For these traits, cluster 1 has the highest mean score. For Cons, the test rejects the hypothesis that clusters 2 and 3 are coming from the same distribution (cluster 3 has the highest mean score and cluster 2 has the lowest). The results on the four-cluster case is in concordance with the three-cluster test results: The tests reject the hypotheses for Ext, Open, and Cons, with cluster 1 having the highest, and cluster 4 having the lowest mean score for Ext and Open. For Cons, Cluster 4 has the highest score. The tests reject the hypotheses for Ext and Open for all comparisons with cluster 4.

TABLE III
STATISTICAL TEST RESULTS ON PERSONALITY IMPRESSION SCORES.

	$H_0 : C_1 = C_2$	$H_0 : C_1 = C_3$	$H_0 : C_2 = C_3$
Ext		X_{t*}^{k**}	X_{t*}^{k***}
Agr			
Cons			X_{t**}^{k**}
Emot			
Open		$X_{t***}^{k\dagger}$	$X_{t***}^{k\dagger}$

X indicates that the null hypothesis, H_0, is rejected;
k: Two-sample K-S test, t: Two-sample t-test;
***: p<<0.001, **: p<0.005, *:p<0.01, †: p<0.05.

These results show that the vloggers in cluster 1 and 2 are on average perceived by zero-acquaintance observers as being more extraverted and open to experience than the vloggers in cluster 3. Cluster 3 contains darker wMEIs in comparison to clusters 1 and 2, emphasizing motion only on the contours of the vloggers' face or upper body, which relates to vlogs that are mainly stationary. This result is in correspondence with previous findings, stating that there is high correlation between the vloggers' activity and extraversion and openness to experience [5]. In contrast, the vloggers in cluster 3 are on average more conscientious than those in cluster 2. Again, based on the properties of the wMEIs, and the corresponding vlogs in cluster 3, this result indicates that conscientiousness people, who are careful, self-disciplined and organized, tend to be more stationary in their vlogs and also edit their videos less. For the four-cluster case, a similar result is obtained. The vloggers in cluster 4 are on average perceived as being less extraverted and open to experience but more conscientious.

Based on the personality impression annotations collected from external observers, we are thus able to show that our clustering approach produces clusters of vlogs that correspond to different vlogging styles and also to vloggers with different personality impressions.

The annotated subset that we use for personality analysis is limited to one vlog per user. We now define one way to project this subset of annotations to the full dataset and present the personality analysis on a larger scale. As people do not change personality in general, we may assume that the personality annotations of each user would be the same for other vlogs of the user. Thus, we propagate the personality impression scores of a user to the other vlogs of that user's video collection, obtaining personality annotations for the whole dataset.

To support the above assumption, we first analyze whether the vlogs in a user's vlog collection are consistent in style, by measuring the percentage of vlogs in a user's vlog collection that are assigned to the same cluster. In our framework each user has between 1 and 8 vlogs, with an average number of 4.8 vlogs. We define a purity measure for each user with respect to the clustering. The measure is defined as the number of videos in the most populated cluster in a user's vlog collection divided by the total number of videos of that user. The purity measure will be 1 if all the vlogs in a users vlog collection is clustered in the same cluster. We excluded the users with only one vlog (49 users) from this analysis, as having a single vlog always leads to a purity measure of 1. Figure 14 shows the distribution of purity with respect to the number of vlogs

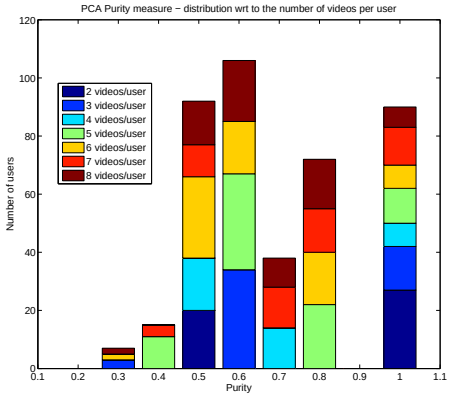


Fig. 14. Cluster purity of users' vlog collections *(Best viewed in color)*

TABLE IV
STATISTICAL TEST RESULTS ON PROJECTED PERSONALITY IMPRESSIONS.

	$H_0 : C_1 = C_2$	$H_0 : C_1 = C_3$	$H_0 : C_2 = C_3$
Ext		X_{t***}^{k**}	X_{t***}^{k***}
Agr			$X_{t\dagger}^{k***}$
Cons			
Emot			
Open		X_{t***}^{k**}	X_{t***}^{k***}

X indicates that the null hypothesis, H_0, is rejected;
k: Two-sample K-S test, t: Two-sample t-test;
***: p<<0.001, **: p<0.005, *:p<0.01, †: p<0.05.

per user for the three-cluster case, for users with more than a single vlog in their collection. The average purity for the whole dataset (for users with more than one vlog) is 0.72. We see that for 90 users (21% of all users with more than one vlog), all the vlogs in their collection is clustered in the same cluster. In comparison to a random clustering result, which corresponds to an average purity measure of 0.56, the purity of our clustering is significantly higher.

Performing the same analysis on the propagated scores, and comparing the results in Table IV to the analysis on the original personality impression scores, we see that the statistical tests reject the same hypotheses for Ext and Open whereas no hypothesis is rejected for Con. In addition, for Agr, the tests reject that clusters 2 and 3 are equal. These results suggest that, even if the propagation of personality is clearly a coarse procedure, the discovered clusters over the full dataset are still meaningful from the perception of social impression.

E. Vlogging Style and Social Attention

As a final way to validate the results of our clustering framework, we focus on the social attention from the perspective of vlogging style, and investigate whether there is a relationship between the social attention that vlogs receive and vlogging style, as indicated by different clusters. As a measure of social attention, we use the log number of views vlogs get. Clearly, there are other measures to characterize attention from audiences but many of them (e.g. number of subscribers) tend to be correlated.

We use the two-sample K-S test to check whether the samples come from the same distribution, where each cluster indicates a sample. We also apply the two-sample t-test to check whether the means of the samples are equal.

TABLE V
MEAN AND MEDIAN LOG-VIEWS OF EACH CLUSTER

	C1	C2	C3
Mean	**5.90**	5.88	5.40
Median	**5.47**	5.36	5.03

TABLE VI
STATISTICAL TEST RESULTS ON THE NUMBER OF VIEWS.

	$H_0 : C_1 = C_2$	$H_0 : C_1 = C_3$	$H_0 : C_2 = C_3$
Log-Views		X_{t**}^{k**}	X_{t***}^{k***}

X indicates that the null hypothesis, H_0, is rejected;
k: Two-sample K-S test, t: Two-sample t-test;
***: p<<0.001, **: p<0.005, *:p<0.01, †: p<0.05.

Table V shows the mean and median log-views for three clusters. Cluster 1 has the highest number of views and cluster 3 has the lowest. The results of the statistical tests based on these are given in Table VI. For three clusters, among the three hypotheses (one for each cluster pair) both tests reject all the hypotheses that involve cluster 3, indicating that cluster 3 is different than the rest of the clusters, having the lowest number of views. For four clusters, the results are in concordance with the previous results with three clusters. Cluster 4, which has the lowest number of views, is found to be significantly different than all other clusters.

In conclusion, the vlogging style represented by cluster 1, which refers to more active and more edited vlogs, receives a significantly higher number of views compared to that of cluster 3 (or cluster 4 when k=4). Cluster 3 represents a vlogging style which is more conversational and stationary, and receives a lower number of views than the other clusters. Our previous studies [6] also show that the social attention is correlated with vlogger's audio-visual activity. In this work, we are able to discover vlogging styles, which correspond to styles differentiated by the vloggers' activity, and there are significant differences between vlogging styles with respect to the number of views they receive. Although in [6] very low correlation ($r = 0.35$, $p > 0.01$) has been found between the level of editing and social attention, in this current work, through the analysis of the discovered clusters, we are able to show that the social attention is related to the vlogging style, which manifests itself in several dimensions, such as the level of editing, the number of shots, the amount of motion, etc.

F. Discussion

Our framework discovered two main vlogging styles. On one end, one style refers to *dynamic* vlogging: the vlogs are highly edited and there is a significant amount of motion either as a result of a physically active vlogger or a moving camera. On the other extreme, another style refers to a "flat" and conversational vlogging: The editing is less and the vlogger is mainly stationary in front of the camera. On top of these physical characteristics, by relating these vlogging styles with personality impressions of the vloggers and with the number of views, we have found significant links between the vlog production and social impressions and

also the social attention they receive. Matching the physical charcteristics of each style with the personality traits linked with them, we see that people scoring higher in extraversion are more active in their vlogs, they edit their videos more, include more non-conversational parts in their vlogs, choose locations with dynamic background, and frame themselves closer to the camera (see Figures 10,12,13 and Tables I,III). These characteristics are also seen in the vlogs of people who score higher with respect to openness to experience. In contrast, people who are percieved as being more intraverted and concientious are less active, edit their videos less and record in more static settings. Our results also indicate that the more active vloggers (who are also found to be more extraverted and open to experience), receive higher number of views than others (see Table VI). Increasing the number of clusters beyond two results in clusters that contain vlogs that are in between these two prototypical vlogging styles. While there are more pronounced differences between two clusters, having more clusters result in the discovery of more nuanced differences in vlogging styles, in between the two extreme cases. Table VII summarizes the properties of the three styles resulting from the three-cluster analysis.

In this paper, we have presented the detailed results of clustering only for PCA features extracted from wMEI. For the other feature sets (Section III-B), we only presented the mean images of clusters (see Figure 5). While it is not possible to present the detailed results for each of the feature sets, mainly for clarity and space reasons, here we include a brief discussion. Although some significant results can be obtained with other feature sets (e.g. HIST), we see that PCA features show higher significance values for the measures that we have presented in this study. For example, while HIST features show significant difference between the clusters for the number of views, they fail to show any difference for the personality traits, or for shot analysis. As also explained in Section V-A, spatial information of the wMEI is important for vlog style discovery and should not be omitted when extracting features.

VI. Conclusions

In this work, we developed a clustering approach to discover a small set of distinct vlogging styles in user-generated web videos based on a spatio-temporal representation of videos.

We use the wMEIs as the spatio-temporal representation of vlogs, which provides a fast and robust way for representing the videos in terms of the dominant motion regions. wMEIs are especially suitable for representing videos that have a default activity in general, i.e. a person sitting in front of the camera, and being computationally efficient, they are suitable to use with large scale data. The wMEI representation could also be used in video retrieval or classification. Exploring the capability of this representation beyond conversational vlogs is an issue for future research.

We have analyzed the discovered vlogging styles in three main dimensions. First, the vlogging styles correspond to different ways of vlog production in terms of physical aspects. One set of vloggers produce vlogs with more motion, more editing, whereas another set of vloggers produce more conversational vlogs with less editing and not much activity. As the second and third dimensions, our analysis shows that these clusters correspond to populations of vloggers that have significantly different scores in several big five personality trait impressions, and whose videos receive significantly different numbers of views, which suggest that the vlogging styles are a reflection of people's characteristics and have a connection to how vloggers are perceived by the YouTube audience.

In order to understand the direct effects of visual information, we have only focused on the visual aspects of vlogs in this work and identified visual vlogging styles. As a future research direction, it would be interesting to look at other channels of information as well, particularly the audio nonverbal channel and the verbal channel, and see whether those channels contribute to the discovery of vlogging styles. Aside from the rich visual information, vlogs contain a variety of audio information, e.g. use of music, use of effects, amount of speech, prosody, etc., which could also be abstracted into vlogging styles. Moreover, the verbal content of a vlog contains significant information that might have an effect on the number of views it would get from the audience. With the analysis of the verbal content, one could investigate the links between the topics of vlogs and the vlogging styles.

References

[1] Amazon Mechanical Turk - https://www.mturk.com.

[2] B. Adams, C. Dorai, and S. Venkatesh. Toward automatic extraction of expressive elements from motion pictures: tempo. *Multimedia, IEEE Transactions on*, 4(4):472 – 481, dec 2002.

[3] B. Barry and G. Davenport. Documenting life: videography and common sense. In *Multimedia and Expo, 2003 (ICME'03), International Conference on*, volume 2, pages 197–200, 2003.

[4] J. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55, 2013.

[5] J.-I. Biel, O. Aran, and D. Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *ICWSM*, 2011.

[6] J.-I. Biel and D. Gatica-Perez. Vlogsense: Conversational behavior and social attention in youtube. *ACM Transactions on Multimedia Computing, Communications and Applications*, 7(1):33:1–33:21, 2011.

[7] J.-I. Biel and D. Gatica-Perez. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proc. of ICWSM*, 2012.

[8] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[9] D. Brezeale and D. J. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems Man and Cybernetics Part C*, 38(3):416–430, 2008.

[10] S. Counts and K. Stecher. Self-presentation of personality during online profile creation. In *AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. CVPR*, volume 2, pages 886–893, 2005.

[12] C. H. Q. Ding and X. He. *K*-means clustering via principal component analysis. In *ICML*, 2004.

[13] W. Gao, Y. Tian, T. Huang, and Q. Yang. Vlogging: A survey of videoblogging technology on the web. *ACM Computing Surveys*, 42(4):15:1–15:57, June 2010.

[14] D. Gatica-Perez, A. C. Loui, and M.-T. Sun. Finding structure in home videos by probabilistic hierarchical clustering. *IEEE Trans. Circuits Syst. Video Techn.*, 13(6):539–548, 2003.

[15] R. Godwin-Jones. Emerging technologies: Digital video update: Youtube, flash, high-definition. *Language Learning & Technology*, 11(1):16–21, 2007.

[16] S. D. Gosling, S. Gaddis, and S. Vazire. Personality impressions based on facebook profiles. In *AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

TABLE VII
SUMMARY OF PROPERTIES OF VLOGGING STYLES

Analysis Category	Cluster 1	Cluster 2	Cluster 3
Shot analysis	More shots Fewer conversational shots Shorter conversational shots Varying shots		Fewer shots More conversational shots Longer conversational shots
Video properties	More motion More dynamic background Larger framing	Higher resolution	Less motion More static background
Personality impressions	More extraversion More openness to experience		Less extraversion More conscientiousness Less openness to experience
Social Attention	More views		Less views

[17] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

[18] M. Griffith. Looking for you: An analysis of video blogs. In *Annual Meeting of the Association for Education in Journalism and Mass Communication*, Washington, DC, 2007.

[19] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:316–322, 2006.

[20] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.

[21] G. Iyengar and A. Lippman. Models for automatic classification of video sequences. In *SPIE Proc. Storage and Retrieval for Image and Video Databases*, pages 216–227, 1997.

[22] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[23] M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing, 7 edition, 2009.

[24] B. Landry and M. Guzdial. Art or circus? characterizing user-created video on youtube. Technical report, Georgia Institute of Technology, 2008.

[25] P. G. Lange. Commenting on comments: Investigating responses to antagonism on YouTube. In *Presented at the Society for Applied Anthropology Conf.*, March 2007.

[26] P. G. Lange. Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, 1(13), 2007.

[27] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[28] H. Molyneaux, S. ODonnell, K. Gibson, and J. Singer. Exploring the gender divide on youtube: An analysis of the creation and reception of vlogs. *American Communication Journal*, 10(1), 2008.

[29] W. Ren, S. Singh, M. Singh, and Y. S. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, 2009.

[30] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 2012.

[31] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 4 edition, 2007.

[32] C. Snoek, M. Worring, and A. G. Hauptmann. Learning rich semantics from news video archives by style analysis. *TOMCCAP*, 2(2):91–108, 2006.

[33] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[34] H. Sundaram and S.-F. Chang. Computable scenes and structures in films. *Multimedia, IEEE Transactions on*, 4(4):482–491, 2002.

[35] L. Wang and D. Suter. Informative shape representations for human action recognition. In *The 18th International Conference on Pattern Recognition (ICPR'06)*, 2006.

[36] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, 2006.

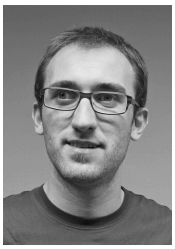
[37] M. Wesch. Youtube and you: Experiences of self-awareness in the context collapse of the recording webcam. *Explorations in Media Ecology*, 8(2):19–34, 2009.

[38] J. Wu and M. Worring. Efficient genre-specific semantic video indexing. *Multimedia, IEEE Transactions on*, 14(2):291–302, 2012.

[39] L. Yang, J. Liu, X. Yang, and X.-S. Hua. Multi-modality web video categorization. *MIR*, page 265, 2007.

[40] C.-C. Yu, H.-Y. Cheng, C.-H. Cheng, and K.-C. Fan. Efficient human action and gait analysis using multiresolution motion energy histogram. *EURASIP J. Adv. Sig. Proc.*, 2010, 2010.

Oya Aran received her PhD degree in Computer Engineering from Bogazici University, Istanbul, Turkey in 2008. She was awarded a EU FP7 Marie Curie IEF fellowship in 2009 and a Swiss National Science Foundation Ambizione fellowship in 2011. Currently, she is a SNSF Ambizione research fellow at the Idiap Research Institute, working on the multimodal analysis of social behavior in small groups. Her research interests include pattern recognition, computer vision, and social computing. She is a member of the IEEE.



Joan-Isaac Biel received his doctoral degree from the Swiss Federal Institute of Technology in Lausanne (EPFL) in June 2013 and is currently a research assistant at the Idiap Research Institute. He has also been visiting researcher at the International Computer Science Institute (Berkeley), HP Labs (Palo Alto), and Yahoo! Labs (Barcelona) working in the areas of audio processing, and social media analysis. His main area of interest is the analysis of social media and social video, with a focus on the personal, social, and communication processes involved on their usage.

Daniel Gatica-Perez (S'01, M'02) is the Head of the Social Computing Group at Idiap Research Institute and Maître d'Enseignement et de Recherche at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. His recent work includes computational methods to understand conversational behavior in social media, urban trends using smartphones and location-based social networks, and emerging social phenomena in face-to-face interaction. Among other professional activities, he has served as Associate Editor of the IEEE Transactions on Multimedia. He is a member of the IEEE.