Visual Interestingness in Webcam-Streams

Helmut Grabner[1,2] Fabian Nater[1,2] Michel Druey[4] Luc Van Gool[1,2,3]

[1]Computer Vision Laboratory
ETH Zurich
{grabner, nater, vangool}@vision.ee.ethz.ch

[2]upicto GmbH
Zurich
{grabner, nater, vangool}@upicto.com

[3]ESAT - PSI / IBBT
K.U. Leuven
luc.vangool@esat.kuleuven.be

[4]Cognitive Psychology Unit
University of Zurich
m.druey@psychologie.uzh.ch

Abstract

Interestingness[1] *is said to be the power of attracting or holding one's attention because something is unusual or exciting. We, as humans, have the great capacity to direct our visual attention and judge the interestingness of a scene. Consider for example the image sequence in the figure on the right. The spider in front of the camera or the snow on the lens are examples of events that deviate from the context since they violate the expectations, and therefore are considered interesting. On the other hand, weather changes or a camera shift, does not considerably raise human attention, even though large regions of the image are influenced. In this work we firstly review related work from psychological, cognitive and computational perspective. Secondly, we investigate what humans consider as "interesting" in image sequences and aggregated the results to a human-consensus-baseline.*

1. Introduction

What do we mean if we find something "interesting"? This frequently used expression is referred to in very broad, often highly subjective terms. Let us consider three examples to illustrate the concept. The scene in which the airplanes crash into the twin towers during the terrorist attack at 9/11, 2001 shockingly shows that such highly unexpected events hugely attract our attention. At the moment when it happened, many of us were fixed to the TV-screens, with a mixture of disbelief, disgust and sympathy for the people in the towers. But up to this day, the images evoke great interest, even after seeing them many times and now knowing exactly what will happen. In contrast, the last minutes of a super-bowl final for example are extremely interesting and capture the attention of many people. This event however gains its attraction from the fact that we don't know how the game will end; but loses relevance to most as soon as the game is over. Finally, human interest is raised from very personal experiences. For example watching the own child playing soccer is much more interesting for the parents than for most others.

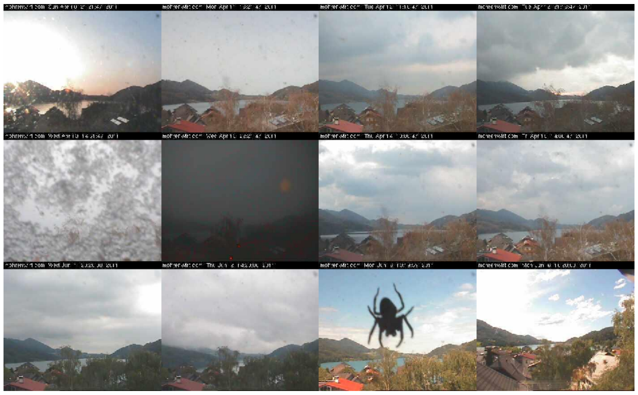


Figure 1. What makes an specific moment in an image stream interesting and how can we computationally approach this question?

These examples illustrate that interestingness highly depends on the context, but also on personal experiences and preferences, which makes it challenging to approach the concept in principled manner (*cf.*, [16]). Humans however have a tremendous capacity to assess how interesting a scene or event is and this greatly helps us to navigate through our daily lives. In order to learn more about human visual perception, but also for commercial purposes (*e.g.*, advertisements), it is of great concern to understand what triggers human attention and interest.

In this work, we restrict ourselves to a particular family of visual input – image sequences recorded by a static video camera – in order to make the problem tractable. In partic-

[1] http://www.thefreedictionary.com/Interestingness, 2012/11/08.

ular, we aim to spot the parts in an image sequence that are considered interesting by many viewers. The relatively constrained setting allows for the discovery of crucial properties of interestingness and will pave the way for further, more challenging scenarios.

2. Related Work

Psychological Perspective. In the mid-50s, Berlyne was among the first to seriously consider interest as psychologically relevant for human learning. In his seminal work [1] he introduced four collative variables, which affect interest: *novelty*, *uncertainty*, *conflict* and *complexity*. More recent research empirically validates and refines this theory, *e.g.*, [3] declare novelty, challenge, attention demand, exploitation intention and instant enjoyment as sources for interestingness (see [16] for a survey and comprehensive discussion). These theories, however, have one limitation as they cannot explain why people respond differently. This is due to the fact that they implicitly trace interest to events rather than to interpretations and appraisals of events. Summarizing, these classical theories have been applied successfully in order to answer the questions what and why some things are interesting to almost everybody. As this is the main focus of the paper at hand, we only take into account these theories.

(Visual) Cognitive Perspective. The concept of interestingness has mostly been studied through understanding which visual stimuli can attract human attention [20]. This is often done by recording gaze patterns of humans watching images or videos (*e.g.*, [5, 7]). Despite a number of unsolved issues, there is common agreement that capturing human attention involves two fundamental properties of human information processing (*cf.* [15]): First, stimulus-based or bottom-up processing and secondly, memory-based or top-down processing. The main bottom-up factors that contribute to the (covert as well as overt) spatial allocation of visual attention are saliency of an object and novelty of an event. The seminal work of Treisman and Gelade [19] introduced the feature-integration theory which has been picked up frequently. For instance, Itti and Koch [9] included three stimulus features (orientation, intensity, and color), and received considerable agreement of their model with human gaze measurements. Motion and abrupt onset are the other features sometimes viewed as relevant in bottom-up processing. Despite the evidence pointing to the crucial role of bottom-up cues, it is obvious that they alone are insufficient to guide visual attention. In fact, it has been repeatedly shown that top-down processes can bias or even override bottom-up visual processing, *e.g.*, [5, 18]. Top-down processing means that individuals are willfully able to track and search for relevant information, while ignoring irrelevant visual stimuli. This processing is strongly affected by task instruction, individual attentional resources, prior knowledge and personal motivation or goals.

Summarizing, it has been shown that saliency and novelty clearly trigger visual attention. However, we argue that this is not necessarily equivalent to interestingness. If a person scans an image or a video in order to understand what is happening, this does not mean that she or he really considers the observations as interesting.

Computational Perspective. Different approaches have been proposed for the automatic detection of visual concepts that relate to interestingness. For example, Johnson and Hogg [11] refer to statistical *outliers* as "possible incidents of interest" or Stauffer and Grimson [17] claim many of their detections to be of "most interest". Other related terms that are often used inconsistently include *surprise*, *saliency*, *abnormality* or *novelty*. The related techniques can be categorized as follows.

Abnormality Detection. In many abnormality detection algorithms, a model of normality is trained from frequent observations. Outliers to these models must be novel concepts and are identified as abnormal events. Typically, such approaches work well for fixed cameras, modeling the entire scene (*e.g.*, [23, 2, 10]) or the behavior of objects within this scene (*e.g.*, [11, 17]). In practice however, it is often unclear to what extent such anomalies are also perceived as interesting by human observers.

Attention Modeling. A simple example shows that "novel" and "interesting" are not always identical. In the white snow paradox, a TV-screen presenting a white noise signal is completely unpredictable and always remains novel, but is unattractive to viewers. This was already noticed in the beginnings of cognitive psychology [1, 22] and calls for a direct modeling of visual concepts which attract human attention. For example, Itti and Baldi came up with a theory of Bayesian surprise [8] or Schmidhuber and coworkers [14, 13] define interestingness as allowing for learning new things: "Neither the arbitrary nor the fully predictable is truly surprising or interesting – only data with still unknown but learnable statistical regularities are".

Interestingness as Category. Machine learning methods are successfully used for many vision tasks, such as object detection or recognition (*e.g.*, detecting faces). Recent works widen these techniques to more complex and less well-defined tasks, for example *emotions* [12], *human memorability* [6] or *aesthetics* [4]. Furthermore, Weinshall *et al.* [21] introduced a novel concept based on disagreement of specific and general classifiers. In all these approaches, a machine learning method is provided with various, low level or specifically designed features. To this end, labeled training data must be available, which is in practice hard to gather in sufficient quality and quantity. For example, the work by Dhar *et al.* [4], use per-image scores from the photo sharing site *flickr*[2]. Yet, we have some doubts about the usefulness of this scores since it is based on properties including clicks, comments or popularity of the photographer that lack of independent human ratings.

3. Human Consensus Base-line

The evaluation of theories and computational techniques always requires a reliable ground truth. In the case of "interestingness" this is clearly a highly subjective judgment and no truly correct answer can be expected from a single person. Therefore, we rely on the knowledge of multiple persons and establish a human consensus that serves as base-line for further investigations.

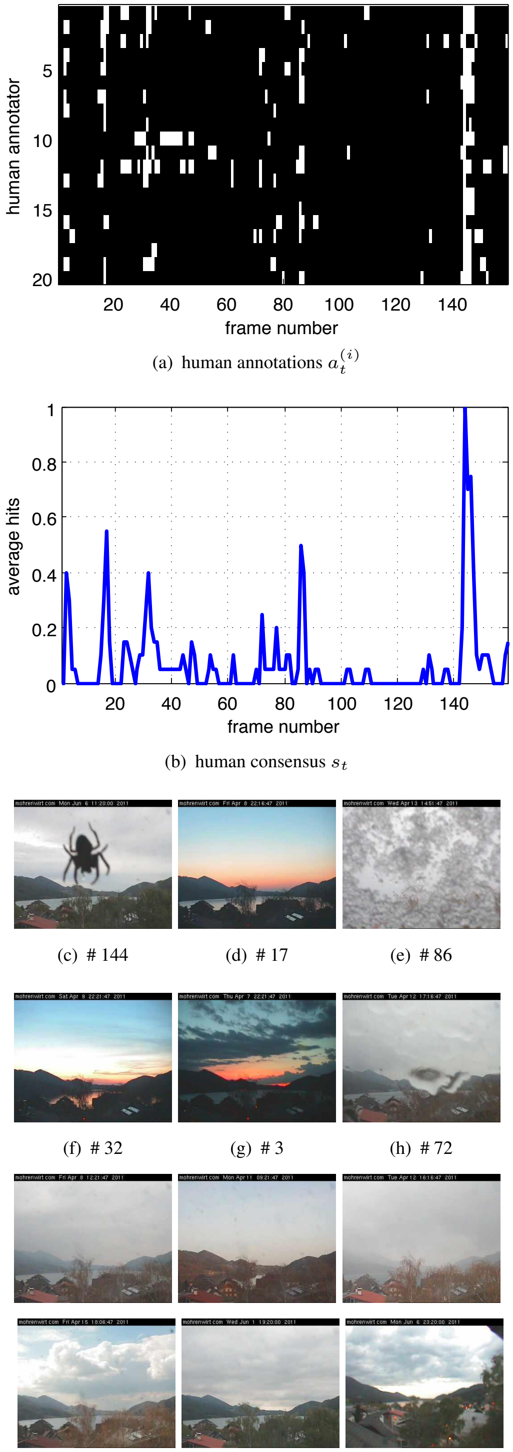
The dataset as well as the established human consensus base-line is available at the authors' webpage or on request.

3.1. Dataset

We chose 20 sequences from publicly available webcams, see Tab. 1. These image sequences were recorded over a long period of time, capturing various – possibly interesting – situations. The clips present typical webcam and surveillance scenes, such as panoramas (*Seq. 1,4,11,17*), highways (*Seq. 5,18*), public squares (*Seq. 3,6,8,15*), urban scenes (*Seq. 10,14,20*), and some particular scenarios (boat rental, *Seq. 2*; stork nest, *Seq. 7*; beach, *Seq. 9*; construction site, *Seq. 12*; the Panama Canal, *Seq. 13*; a port, *Seq. 16*; and the Tower Bridge, *Seq. 19*). Image resolutions range from 352 × 288 (PAL) up to 420 × 315. Images were recorded during several days and usually sampled at one frame per hour. We manually selected representative sub-sequences, *e.g.*, excluding very dark night images. The finally displayed image sequence consisted of 159 color images, continuously displayed at approximately 1 fps.

3.2. User Study

Setup. 26 male and 20 female test persons, aged between 18 and 47 and having normal or corrected vision participated in the test. They were instructed to watch the image sequence, press a button if they considered something as interesting, and they had to press the button again to release the interestingness tag. No further instruction was given and hence the participants were free to judge what they considered as interesting. Furthermore, they were asked to rate the overall interestingness of every clip at the end of the experiment, once they had seen all image sequences. In order to recall the sequences, the beginning of each clip was replayed briefly, and ratings were asked in 7 levels (1 = very boring, 7 = very interesting). Test persons completed the



(a) human annotations $a_t^{(i)}$

(b) human consensus s_t

(c) # 144 (d) # 17 (e) # 86

(f) # 32 (g) # 3 (h) # 72

(i) uninteresting (score = 0.00)

Figure 2. Exemplary user annotations (a) and the obtained human consensus for *Seq. 1* (b). Interesting parts (c)-(h) include abnormal, unexpected events as well as "aesthetic" images. Images consistently considered as uninteresting (i) might have large variations on pixel level (*e.g.*, the sky region) but are semantically "normal".

[2] `http://www.flickr.com/explore/interesting/`, 2012/11/12.

task unwatched and they could stop when they reached their personal time budget. Sequences and their ordering were chosen randomly. Each participant evaluated between 6 and 10 sequences and all sequences were viewed by at least 20 persons.

Human Consensus. Let $a_t^{(i)}$ be the individual binary interestingness annotations of person i for image I_t for a particular sequence. If the user considers the frame as interesting $a_t^{(i)} = 1$, and 0 otherwise. The human consensus interestingness score is defined as the per-frame average of the individual annotations for a particular sequence, *i.e.*, $s_t = \frac{1}{N} \sum_i a_t^{(i)}$, were N is the number of participants who annotated this sequence. Similarly, the overall interestingness rating r of a sequence is the average of the individuals' overall ratings. If many individuals consider a frame interesting (*i.e.*, $s_t > 0.5$), we call this an interesting event. Hence, the most interesting events in a sequence can be ranked with respect to their interestingness score, that is the agreement among individuals.

Example Sequence. A detailed example for the first sequence in the data-set is given in Fig. 2, showing the raw user annotations (a), the average across persons – the established human consensus – (b) and examples of highly interesting (c)-(h) and uninteresting (i) frames. Interesting frames include surprising and unexpected events (the spider, snow and rain on the lens) as well as aesthetic images (sunsets). Note that many other frames also might show large variations (especially in the sky region) but are consistently considered as "normal" or "uninteresting". Furthermore, at the very end of this sequence the camera was moved quite essentially (zoomed in). This change however, was mostly ignored by the viewers. Hence, it seems that abstraction and semantic interpretation of what we expect is essential.

Dataset Overview. Tab. 1 summarizes results of the human responses. For each sequence are shown: a typical, (uninteresting) image, the most interesting image with its corresponding score s_{max}, the number of interesting and uninteresting events, as well as an overall interestingness rating. Some events clearly attract the focus of many viewers, *e.g.*, in *Seq. 6*, many things are happening in the observed square (market, auto show, sports game, *etc*.). In each clip however, there are frequent intervals that are consistently labeled as uninteresting.

3.3. Consistency

Per-frame Score. In order to quantify the consistency within the responses of test persons, we use standardized Cronbach's alpha. This widely used measure for the reliability of a psychometric test is defined as $\alpha_{st} = \frac{n\bar{r}}{1+(n-1)\bar{r}}$, where n is the number of persons and $\bar{r}$ the mean correlation between each of them.[3] As can be seen from Tab. 1, the measured consistency is generally high for most sequences (avg. $\bar{\alpha_{st}} = 0.83$, max. $\alpha_{st,max} = 0.93$). The responses of some sequences are less consistent (*e.g.*, *Seq. 11*, $\alpha_{st} = 0.75$), this is due to the influence of individual preferences such as special cloud formations or sunsets.

Overall Rating. We use Sperman's rank coefficient ρ to assess the consistency across the participants overall ratings of the viewed sequences. This measure reflects how well two variables can be explained by a monotonic relation, therefore we first rank the annotated scenes according to the users' ratings. Overall we achieved a mean rank coefficient of $\bar{\rho} = 0.30$ across all participants ($\rho_{max} = 0.94$; $\rho_{min} = -0.83$). This rating consistency is relatively weak, compared to the consistent annotations of per-frame interestingness. Remarkably, we spot an a more significant correlation $\bar{\rho}_{\mu_s} = 0.41$ between the average interestingness score μ_s of a sequence with its' ranking. Hence, having reliable interestingness scores per frame would also allow for a rough overall ranking.

Summarizing, the human consensus base-line of per-frame scores gives a solid ground and can be used to (i) build computational models and (ii) evaluate them.

4. Conclusion

We recorded moments of human interest and aggregated them to a consensus. Apparently, the image sequence specifies the context in which humans judge events as interesting. Hence, outlier detection techniques permit to capture a substantial fraction of interestingness already. While many abnormal events are consistently considered to be interesting, also a large portion of them are not, such as camera failures or different cloud formations. On the other hand, statistically well-explained, normal events might be still interesting, *e.g.*, raising of Tower Bridge.

This work is a first attempt to quantify visual interestingness. It will be interesting to investigate on (1) building and comparing computational models for visual interestingness; (2) examine the relationship between visual interestingness and other measures, such as memorability or image quality; (3) widen the scope to more general settings; and (4) taking the specific preferences of a particular observer into account.

References

[1] D. Berlyne. *Conflict, arousal, and curiosity*. McGraw-Hill, 1960.

[2] M. Breitenstein, H. Grabner, and L. Van Gool. Hunting nessie: Real time abnormality detection from webcams. In *Proc. IEEE WS on Visual Surveillance*, 2009.

[3] Literature suggests, $\alpha_{st} > 0.7$: acceptable; $\alpha_{st} > 0.8$: good; and $\alpha_{st} > 0.9$: excellent.

[3] A. Chen, P. Darst, and R. Pangrazi. An examination of situational interest and its sources. *Brit. J. of Edu. Psychology*, 71:383–400, 2001.

[4] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proc. CVPR*, 2011.

[5] J. Henderson. Human gaze control during real-world scene perception. *TRENDS in Cognitive Science*, 7(11):498–504, 2003.

[6] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proc. CVPR*, 2011.

[7] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. CVPR*, 2005.

[8] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49:1295–1306, 2009.

[9] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[10] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *Proc. CVPR*, 2007.

[11] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *Proc. BMVC*, 1995.

[12] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proc. ACM Multimedia*, 2010.

[13] T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber. Coherence progress: A measure of interestingness based on fixed compressors. In *In Proc. Conf. on Artificial Gen. Intelligence*, 2011.

[14] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. Technical report, TU-Munich, 2009.

[15] W. Schneider and R. Shiffrin. Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review*, 84:1–66, 1977.

[16] P. Silvia. *Exploring the psychology of interest*. Oxford University Press, 2006.

[17] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–757, 2000.

[18] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.

[19] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[20] J. Tsotsos, L. Itti, and G. Rees. A brief and selective history of attention. In *Neurobiology of Attention*. Elsevier, 2005.

[21] D. Weinshall, A. Zweig, H. Hermansky, S. Kombrink, F. Ohl, J. Anemüller, J. Bach, L. Van Gool, F. Nater, T. Pajdla, M. Havlena, and M. Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *PAMI*, 2012.

[22] W. Wundt. *Grundzüge der physiologischen Psychologie*. Engelmann, Leipzig, 1874.

[23] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. CVPR*, 2004.

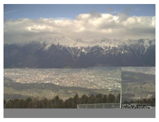
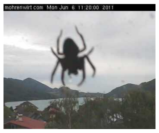
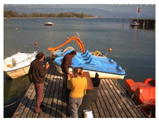
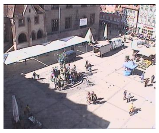
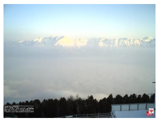
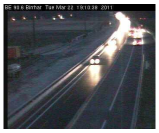
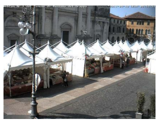
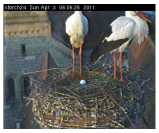
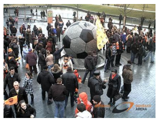
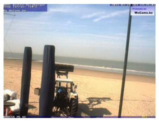
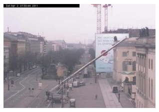
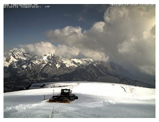
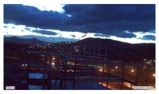
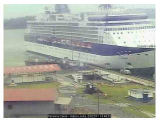
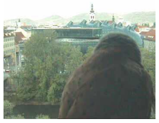
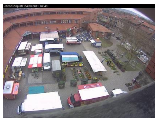
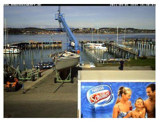
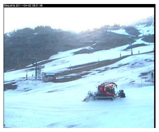
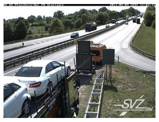
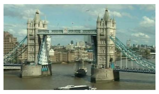
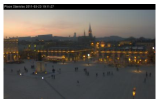
Seq.	1	2	3	4	5	6	7
typical image							
most interesting image							
s_{max}	1.00	0.75	0.50	0.75	0.45	0.95	0.70
$s(\mu \pm \sigma)$	0.07 ± 0.14	0.14 ± 0.15	0.10 ± 0.11	0.12 ± 0.14	0.06 ± 0.08	0.21 ± 0.25	0.19 ± 0.14
$\#s > 0.50$	4	5	0	3	0	26	7
$\#s < 0.25$	146	123	136	129	152	100	113
α_{st}	0.90	0.82	0.73	0.79	0.74	0.93	0.82
$r(\mu \pm \sigma)$	2.5 ± 1.2	3.6 ± 1.0	2.5 ± 1.1	3.4 ± 1.5	1.9 ± 0.8	4.0 ± 1.3	4.7 ± 1.7
Seq.	8	9	10	11	12	13	14
typical image							
most interesting image							
s_{max}	0.80	0.70	0.95	0.70	0.70	0.80	0.80
$s(\mu \pm \sigma)$	0.05 ± 0.11	0.07 ± 0.13	0.08 ± 0.17	0.13 ± 0.14	0.09 ± 0.13	0.16 ± 0.18	0.11 ± 0.17
$\#s > 0.50$	2	4	8	2	3	7	7
$\#s < 0.25$	152	143	139	125	144	114	135
α_{st}	0.87	0.88	0.92	0.75	0.80	0.85	0.88
$r(\mu \pm \sigma)$	2.7 ± 1.1	2.5 ± 1.1	2.8 ± 1.1	4.0 ± 1.3	2.2 ± 1.2	3.6 ± 1.1	2.8 ± 1.3
Seq.	15	16	17	18	19	20	average
typical image							
most interesting image							
s_{max}	0.65	0.80	0.65	0.55	0.70	0.80	0.74
$s(\mu \pm \sigma)$	0.13 ± 0.18	0.13 ± 0.15	0.12 ± 0.14	0.06 ± 0.10	0.11 ± 0.14	0.09 ± 0.13	0.11
$\#s > 0.50$	11	6	5	1	4	3	5
$\#s < 0.25$	126	124	134	147	135	140	133
α_{st}	0.86	0.82	0.78	0.78	0.83	0.82	0.83
$r(\mu \pm \sigma)$	2.9 ± 1.0	3.2 ± 1.4	2.9 ± 1.6	1.9 ± 0.7	3.5 ± 1.6	3.6 ± 1.6	3.0

Table 1. Dataset and statistics of the obtained human consensus base-line. **Available at the authors' webpage or on request.**