SOCIAL FOCUS OF ATTENTION AS A TIME FUNCTION DERIVED FROM MULTIMODAL SIGNALS

Danil Korchagin[1] *and Hamid Reza Abutalebi*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Speech Processing Research Lab (SPRL), Elec. and Comp. Eng. Dept., Yazd University, Yazd, Iran

ABSTRACT

In this paper, we present the results of a study on the social focus of attention as a time function derived from the multisource multimodal signals, recorded by different personal capturing devices during social events. The core of the approach is based on fission and fusion of multichannel audio, video and social modalities to derive the social focus of attention. The results achieved to date on 16+ hours of real-life data prove the feasibility of the approach.

Index Terms — Multimodal signal processing, data analysis, sensor fusion

1. INTRODUCTION

The TA2 (Together Anywhere, Together Anytime) project [1] is concerned with investigation of how multimedia devices can be introduced into a family scenario to break down technology and distance barriers. Technically, the TA2 project tries to improve group-to-group communication by making it more natural and by giving the users the means to easily participate in shared activities. In this sense, we are interested in the use of consumer level multimedia devices in novel application scenarios.

One generic scenario is the use of multiple capture devices at the same social event (see Fig. 1). The primary characteristic of a social event is that the focus of attention of the group of attendees is dedicated to the most important/interesting moments of the event and vice-versa. Today, with the ease of media content migration, millions of people share with others their media assets, recorded at social events. This phenomenon, known as crowdsourcing, is exemplified by the web sites such as Facebook, YouTube, and others. In this sense the social focus of attention can be considered as a feedback-based validation [2, 3] (also known as a popularity measure) of shared media content within large communities or as a social network analysis [4] of the group of attendees. In our study we concentrate mainly on the ability of dealing effectively with social interactions during the social events to capture the attention of the audience from multisource multimodal media assets.

Fig. 1. An example of social event media coverage. The event consists of a music performance with 190 media assets from 12 cameras/people. The recordings are automatically plotted versus time (in seconds), taking the lowest stream with available time slot.

Social signal processing is a recent domain aiming at bringing social intelligence to computers [5]. The good survey on social signals and their function can be found in [6]. The study [7] on automatic analysis of conversational vlogs proposes to measure social attention by the number of views. Other related studies rely on video modality [8, 9], joint audio-visual modalities [10] or joint video-contextual modalities [9], where the social focus of attention is often called as shared focus or joint focus of attention. Typically, the corresponding techniques rely on assumption of static cameras and impose controlled environments.

In the context, TA2 presents several challenges: the environments are unconstrained; the devices are hand-held and are turned on and off at the will of their users; the recordings captured at the same time by different cameras may look and sound different; the corresponding metadata is divergent; the capture devices are neither calibrated, nor synchronised. If we could prove the feasibility of automatically derived attention of the audience from multisource multimodal media assets, then higher level routine automation (e.g., multimedia clustering, interpretation of social sensing, cut-point suitability estimation and authoring processes [11]) could benefit from additional high-level semantic information.

2. SOCIAL FOCUS OF ATTENTION

We define the social focus of attention as a relative popularity measure dedicated to a shared target in the audiovisual domain within social group of people ("how much audiovisual attention is attracted by the target"). Let $M=\{m_i\}$ be an open set of all available media assets, recorded by social group of people across different events $E=\{E_j\}$. Each media asset m_i can be represented as a couple (a_i,v_i), where a_i is an audio stream and v_i is a video stream, captured concurrently by device. Then the social focus of attention $sfoa_j(t)$ from the social event E_j at time instant t can be defined as a proportion of media assets with correlated audio and visual focus of attention at each time instant within the specified event:

$$sfoa_j(t)=\frac{size\left(\{m_i\}\middle|\begin{array}{c} m_i(t)\in E_j \\ length(m_i)\geq 2s \\ vfoa(m_i(t))=afoa(m_i(t)) \end{array}\right)}{\max\limits_{\tau}\left(size\left(\{m_i\}\middle|\begin{array}{c} m_i(\tau)\in E_j \\ length(m_i)\geq 2s \end{array}\right)\right)}$$

In the above equation, *size* is size of the set in number of media assets, $vfoa(m_i(t))$ is visual focus of attention and $afoa(m_i(t))$ is audio focus of attention of the media asset $m_i(t)$ at time instant t. The condition $length(m_i)\geq 2s$ is used to eliminate accidental short-term recordings (less than 2 seconds long).

The condition $m_i(t)\in E_j$ can be verified via out-of-scene data detection [12], which in turn is based on synchronisation confidence estimation [13]. Corresponding synchronisation and confidence estimation, based on the time-quefrency signatures, are performed by searching for a best distance in n-dimensional Euclidean space between the time-quefrency representations A_i and A_j^r of the test and reference audio signals a_i and a_j^r [13]. The reference audio signal a_j^r is taken from a camera that recorded the whole event E_j without interruptions. It is used only for the synchronisation purpose and can be eliminated, if it is not required by the involved synchronisation mechanism.

The relative position within the reference signal a_j^r from social event E_j is given by:

$$t_j^i=\alpha\cdot\arg\min_{A_j^r}(d(A_i,A_j^r))$$

where d is Euclidean metric, α is a step within time-quefrency representation in s.

The condition $m_i(t)\in E_j$ can be rewritten using a confidence estimation [12] as a measure of relative variance of the search space via minimum and maximum distances:

$$\frac{\left|\mathrm{E}(d(A_i,A_j^r))-\min\limits_{A_j^r}(d(A_i,A_j^r))\right|}{4\cdot\left|\mathrm{E}(d(A_i,A_j^r))-\max\limits_{A_j^r}(d(A_i,A_j^r))\right|}-0.2l_{(i)}^{-1}\geq c$$

In the above equation, c is the confidence threshold for successful synchronisation of the test (a_i) and reference (a_j^r) signals. E is expectation. $l_{(i)}$ is the length of the test signal a_i in seconds.

Other audio-based solutions are based on the fast cross correlation of the signals, audio onsets [14], or audio fingerprinting techniques [15-18]. Most of them results in fairly good but not perfect synchronisation of the recordings in real-life conditions.

The condition $vfoa(m_i(t))=afoa(m_i(t))$ can be approximated by $doa(v_i(t))=doa(a_i(t))$, where *doa* is the estimated direction of arrival. The direction of arrival of sound (to the stereo microphone array) can be represented as an angle with respect to some reference direction (0°). We define this reference direction as an imaginary arrow intersecting the consumer level device at the centre of the stereo microphone array, facing the video scene. Taking into account that within all consumer level devices video sensor is located in the parallel surface in respect to the stereo microphone array, the corresponding condition can be replaced by the constraint on time delay of arrival, which can be estimated based on well known Generalized Cross Correlation (GCC) [19].

Generalized cross correlation with maximum likelihood weighting (GCC-ML) is theoretically optimal in the presence of uncorrelated noise, nevertheless its performance degrades with increasing reverberation [20]. In addition, it requires the spectral information of the noise from the preceding noise-only frames of the stereo microphone array, which usually cannot be reliably achieved during noisy social events. Generalized cross correlation with phase transform weighting (GCC-PHAT) is more robust against reverberation [21] due to the whitening of the microphone array signals. Also, it does not require any information about precedent noise levels. Therefore, the condition $vfoa(m_i(t))=afoa(m_i(t))$ can be finally approximated by:

$$\left|\arg\max\left(F^{-1}\left(\frac{F\{a_i^{(l)}(t)\}\cdot\left(F\{a_i^{(r)}(t)\}\right)^*}{\left|F\{a_i^{(l)}(t)\}\cdot\left(F\{a_i^{(r)}(t)\}\right)^*\right|}\right)\right)\right|\leq\tau_{foa}$$

In the above equation, F denotes the Fourier transform. An asterisk indicates the complex conjugate. $a_i^{(l)}(t)$ and $a_i^{(r)}(t)$ are the left and right channels of the audio stream $a_i(t)$ within the media asset $m_i(t)$. τ_{foa} is the threshold for the corresponding time delay of arrival.

There is a trade-off between GCC-PHAT robustness and time resolution. While a long analysis window leads to a reduction in the time resolution, a short analysis window reduces the robustness of the corresponding cross correlation and in turn results in unreliable estimations of the time delay of arrival. Recent results on GCC-PHAT technique studies can be found in [22-24].

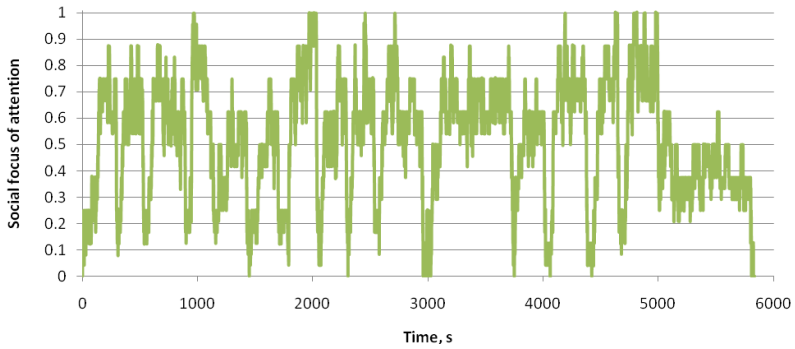

Fig. 2. An example of social focus of attention as a time function.

3. EXPERIMENTIAL RESULTS

The results presented in this paper were achieved on a real life dataset of 508 recordings:

- 4 reference signals (total length – 3 h 47 min), recorded with:
  - Canon XL-G1,
  - 3x Sony HDR-520VE.
- 504 test signals (97, 79, 139, 189 test signals per corresponding reference signal; total length – 12 h 21 min), recorded with:
  - Canon HD-HSF10, Powershot S5IS, FS100E mini, XM1 mini DV,
  - iPhone 3G S,
  - Nikon D70,
  - Nokia N95,
  - Panasonic Lumix DMC-F57, DMC-FX500, DMC-LX3,
  - Sanyo Xacti HD mini,
  - Sony DCR-PC3e, PDC-10E, PDC-100E,
  - etc.

The recordings were captured by several social groups of people (with up to 12 socially connected people per group) during 4 different events in 2 different countries. The reference signal contents consist of musical concerts/rehearsals with multiple sub-events/replays one after the other. No constraints were applied for the test recordings. The corresponding devices were turned on and off at the will of their users. Though the described dataset contains only musical events, the proposed technique is applicable to other types of social events as well.

Experiments were conducted on an open set, which resulted in 2016 possible combinations for the condition $m_i(t) \in E_j$. All corresponding audio tracks were extracted and converted to 48 kHz stereo PCM files with FFMPEG software [25]. Corresponding time delays of arrival were calculated per each recording in step of 1/3 s. The condition $vfoa(m_i(t))=afoa(m_i(t))$ was triggered in 93.2% of the cases, although the corresponding performance levels were not ideal. Finally, the results were averaged in step of 1 s to derive the social focus of attention as a time function per each event.

Fig. 2 shows the social focus of attention versus time for the event illustrated in Fig. 1. The valleys on the graph correspond to the lower attention due to transitions between sub-events (e.g. change of performers, pause between songs). While for the events with a high (≥ 10) number of involved personal devices it was always evident to locate sub-events, for the events with fewer devices, the estimated social focus of attention was sometimes ambiguous.

In Fig. 3 we illustrate the dependency between precision and recall values for the condition $m_i(t) \in E_j$. Precision is defined as the number of true positive test signals (test signals correctly detected as belonging to the positive class) divided by the total number of test signals detected as belonging to the positive class (the sum of true positive and false positive test segments). Recall is defined as the number of true positive test signals divided by the total number of test signals that actually belong to the positive class (the sum of true positive and false negative test signals). It is clearly visible, that the applied time-quefrency signature based technique outperforms well known cross correlation (which is given for comparison purpose only). We were able to achieve 100% precision in the case of 98.4% recall, while the cross correlation resulted in 90% precision in the case of 90% recall.

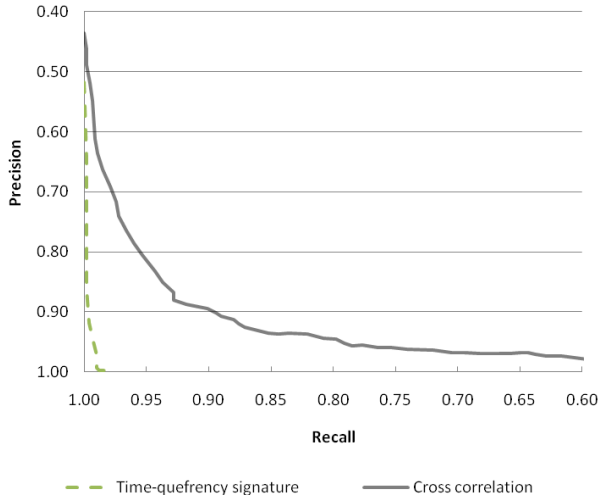

Fig. 3. Precision versus recall for signal clustering.

4. CONCLUSION

We have shown the feasibility of automatic derivation of the social focus of attention from multisource multimodal signals, recorded by different personal capturing devices during social events. We found that the social focus of attention can be inferred from relations between audio, visual and personal focus of attention across crowdsourced media assets. Performance levels achieved to date on 16+ hours of real-life dataset have shown sufficient reliability. The achieved results are promising for the further development of the concept in several directions such as improvement of relative direction of arrival estimation, experiments on datasets with higher level of media asset density and investigations on its application in the subsequent higher level components.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project “Together Anywhere, Together Anytime” (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793. We are grateful to British Telecom and Centrum Wiskunde & Informatica for provision of the real life dataset.

6. REFERENCES

[1] Integrating project within the European research programme 7, "Together anywhere, together anytime", http://www.ta2-project.eu, 2008.

[2] B. A. Huberman, "Crowdsourcing and attention", Computer, vol. 41, issue 11, pp. 103–105, 2008.

[3] J. Burgess and J. Green, "YouTube: Online video and participatory culture", Polity Press, Cambridge, UK, 2009.

[4] S. Favre, "Social Network Analysis in Multimedia Indexing: Making Sense of People in Multiparty Recordings", Proc. of the International Conference on Affective Computing & Intelligent Interaction (ACII), Enschede, Netherlands, 2009.

[5] A. Pentland, "Social signal processing", IEEE Signal Processing Magazine, 24(4): 108–111, 2007.

[6] A. Vinciarelli, M. Pantic, H. Bourlard and A. Pentland, "Social Signals, their Function, and Automatic Analysis: A Survey", Proc. of International Conference on Multimodal Interfaces, 2008.

[7] J.-I. Biel and D. Gatica-Perez, "Vlogcast Yourself: Nonverbal Behavior and Attention in Social Media", Proc. International Conference on Multimodal Interfaces, 2010.

[8] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues", IEEE Trans. on Neural Networks, vol. 13(4), pp. 928–938, 2002.

[9] S. O. Ba and J.-M. Odobez, "Multi-Person Visual Focus of Attention from Head Pose and Meeting Contextual Cues", IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(1): 101–116, 2011.

[10] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A probabilistic inference of multiparty-conversation structure based on Markov switching models of gaze patterns, head directions, and utterances", Proc. of the International Conference on Multimodal Interfaces, pp. 191–198, 2005.

[11] P. Cesar, D. Geerts, and K. Chorianopoulos, "Social Interactive Television: Immersive Shared Experiences and Perspectives", Information Science Reference, Hershey, New York, 2009.

[12] D. Korchagin, "Out-of-scene AV data detection", Proc. IADIS International Conference on Applied Computing, vol. 2, pp. 244-248, Rome, Italy, 2009.

[13] D. Korchagin, P. N. Garner, and J. Dines, "Automatic temporal alignment of AV data with confidence estimation", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, USA, 2010.

[14] J. P. Bello, L. Daudet, S. Abdallah, et al., "A tutorial on onset detection in music signals", IEEE Trans. on Speech and Audio Processing, vol. 13, issue 5, part 2, 2005.

[15] M. Cremer and R. Cook, "Machine-assisted editing of user generated content", Proc. of SPIE-IS&T Electronic Imaging, vol. 7254, 2009.

[16] L. Kennedy and M. Naaman, "Less talk, more rock: automated organization of community-contributed collections of concert videos", Proc. of the 18th ACM International Conference on World Wide Web, pp. 311–320, 2009.

[17] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system", Proc. of the International Symposium on Music Information Retrieval, 2002.

[18] P. Shrestha, H. Weda, and M. Barbieri, "Synchronization of multi-camera video recordings based on audio", Proc. of the 15th annual ACM International Conference on Multimedia, pp. 545–548, 2007.

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 24(4), pp. 320–327, 1976.

[20] B. Champagne, S. Bedard and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation", IEEE Trans. on Speech Audio Processing, vol. 4(2), pp. 148–152, 1996.

[21] C. Zhang, D. Florelncio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2565–2568, 2008.

[22] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview", EURASIP Journal on Applied Signal Processing, vol. 2006, id 26503, 2006.

[23] K. Byoungho, P. Youngjin, P. Younsik, "Analysis of the GCC-PHAT technique for multiple sources", Proc. of the International Conference on Control Automation and Systems (ICCAS), pp. 2070–2073, 2010.

[24] Q. Bo, Z. Heng, F. Qiang, Y. Yonghong, "Subsample time delay estimation via improved GCC-PHAT algorithm", Proc. of the International Conference on Signal Processing (ICSP), pp. 2579–2582, 2008.

[25] Open source multiformat multimedia conversion tool "FFMPEG", http://www.ffmpeg.org.