AUTOMATIC MATCHING AND SYNCHRONIZATION OF USER GENERATED VIDEOS FROM A LARGE SCALE SPORT EVENT

Nikolaos Stefanakis,[1] *Stavros Chonianakis* [2] *and Athanasios Mouchtaris* [1,2]

[1]Foundation for Research and Technology - Hellas, Institute of Computer Science, 70013 Heraklion, Crete, Greece
[2]University of Crete, Department of Computer Science, 70013 Heraklion, Crete, Greece

ABSTRACT

Exploiting correlations in the audio, several works in the past have demonstrated the ability to automatically match and synchronize User Generated Video (UGV) files of the same event. In this paper, we focus on the challenging acoustic environment of a large scale athletic event. We show that the chanting of the crowd produces an acoustic background common in the audio streams of different UGVs and we design a novel audio fingerprinting method for organizing the UGV collection based on that content. Results presented with recordings from a crowded football match demonstrate that the proposed approach provides significantly better audio matching performance in comparison to three of the most well known audio fingerprinting techniques.

***Index Terms*—** audio matching, audio synchronization, audio fingerprinting, content based management, user generated content

1. INTRODUCTION

With the proliferation of smart-phones and portable electronic devices, more and more of us become engaged in the process of capturing and sharing audiovisual content from public events that we attend. Such content can be very valuable to the broadcasters and producers as it may enrich the professional footage or provide coverage for parts of the event that have not been captured by the professional equipment. Yet, it is not trivial to organize this content in a way that it can be usable for the said purpose. For example, as user generated content lacks metadata which is informative about the exact location and time of recording, it would require enormous time and effort from the professional editor to manually search for videos referring to a particular segment of the event, or to group and temporally align multiple videos overlapping in space and time.

Fortunately, as several works have demonstrated in the past, it is possible to automatically organize such content by exploiting the correlations in the audio streams available in the UGV collection. Of fundamental importance for succeeding in this task is to employ audio features which are robust to different kinds of channel variations – as for example environmental noise, different recording locations, varying frequency response of each recording device, etc – and which at the same time allow for fast mining through large collections containing hundreds or thousands of UGVs. Respecting these requirement, audio fingerprinting techniques, originally designed for the purpose of song identification [1–3], have been successfully employed from researchers working on the management of user generated content in the past [4–8].

In this paper, our goal is to quantify the ability of different fingerprinting techniques in audio matching and synchronization, i.e. the process of identifying audio (and as a consequence video) recordings overlapping in time, as well as the time-offsets which are required for their time-alignment. Being successful in this task is an important prerequisite for various applications, as for example, to produce a linear plot with multiple visual perspectives of the same event [5], or to combine the audio streams from the different devices in order to produce a new acoustic sequence of increased duration [8] and enhanced quality [9, 10]. In this work, we focus on the case of a large scale athletic event and in particular, a football match taking place inside a crowded open stadium. We show that the chanting of the crowd produces a distinct acoustic background common to the generated audio streams, and we propose a novel acoustic feature which is tailored for correlating with this particular type of content. Results presented with recordings from a real football match demonstrate that the proposed approach provides significantly better performance in comparison to three of the most well known audio fingerprinting techniques.

2. METHODOLOGY

To find instances of the same event across a collection of M video recordings, we perform three basic steps. First, the audio is imported from each video file and stored at a separate folder in standard PCM format. For each audio file, an $N_i \times B$ audio fingerprint matrix $\mathbf{F}_i$ is constructed and stored, where N_i is the number of time-frames used in the analysis of the

ith recording while B depends on the dimensionality of each audio fingerprinting technique. Third, all the $M(M-1)/2$ pairwise combinations of fingerprints are presented as input to an audio matching algorithm. For each pair ij, the process provides us with a vector containing the values of a generalized cross-correlation function denoted as $R_{\mathbf{F}_i,\mathbf{F}_j}(\tau)$, where $\tau \in \mathbb{Z}$ spans all possible time-frame offsets between fingerprints i and j. Similar as in [8], we use the maximum of the cross-correlation

$$p_{ij} = \max_{\tau} R_{\mathbf{F}_i,\mathbf{F}_j}(\tau), \tag{1}$$

as the confidence metric for deciding whether recordings i and j should be paired or not. A side product of this process is the time-frame difference of arrival $\hat{\tau}_{ij} = \arg\max_{\tau} R_{\mathbf{F}_i,\mathbf{F}_j}(\tau)$ which potentially synchronizes recordings i and j. A pair is assigned a "positive match" if p_{ij}, the so-called match strength from now on, is equal or greater to the value of a predefined threshold θ and a "negative match" in the opposite case. The exact way that function $R_{\mathbf{F}_i,\mathbf{F}_j}(\tau)$ is defined varies in accordance to the nature of each fingerprinting technique.

For assessing the audio matching performance, we define two simple metrics, namely, the Positive Match Score (PMS) and the Negative Match Score (NMS). The PMS [resp. NMS] is defined as PMS $= \frac{Q^+}{N^+}$ [resp. NMS $= \frac{Q^-}{N^-}$] where Q^+ [resp. Q^-] is the number of pairs correctly assigned a positive [resp. negative] match and N^+ [resp. N^-] is the number of pairs which should have been assigned a positive [resp. negative] match. Obviously, $N^+ + N^- = M(M-2)/2$ holds. Naturally, the value of the threshold used within each approach needs to guarantee an optimal trade off between PMS and NMS, as a large [resp. small] value of the threshold will improve NMS [resp. PMS] in the cost of a low PMS [resp. NMS].

2.1. Philips Robust Hash (PRH)

The first approach that we consider for audio fingerprinting is a slightly modified version of the method presented in [2], latter exploited for audio organization by the authors in [4, 11]. The technique takes into account the sign of the energy differences simultaneously along the time and frequency axes. The spectrum is divided into B non-overlapping subbands and the widths of subsequent subbands increase in a logarithmic way with frequency. Let $Z_i(n,b)$ symbolize the energy at the nth time-frame and bth subband region. If we define $\Delta Z_i(n,b) = Z_i(n,b) - Z_i(n,b+1) - (Z_i(n-1,b) - Z_i(n-1,b+1))$ then, the pixel value that corresponds to the bth subband at the nt time-frame can be determined based on the following equation:

$$F_i(n,b) = \begin{cases} 1 \text{ if} & \Delta Z_i(n,b) > 0 \\ \text{-1 if} & \Delta Z_i(n,b) \leq 0. \end{cases} \tag{2}$$

A time-offset and a matching decision for two recordings i and j is obtained using the process described in the beginning of section 2, with the generalized cross-correlation function defined as

$$R_{\mathbf{F}_i,\mathbf{F}_j}(\tau) = \sum_{n=-\infty}^{\infty} \mathbf{F}_i^T(n)\mathbf{F}_j(n+\tau), \tag{3}$$

where $(\cdot)^T$ denotes matrix transposition and $\mathbf{F}_i(n)$ is the $B \times 1$ feature vector specific to time-frame n. Use of the binary set $\{-1, 1\}$ in Eq. (2) instead of the set $\{0, 1\}$, originally proposed in [2], allows for implementation with a fast cross-correlation algorithm such as Matlab function **xcorr**.

2.2. Audio chroma based fingerprinting

Correlating strongly to the harmonic information contained in the audio signals, audio chroma is a well known feature for content-based audio analysis [7, 12, 13]. Essentially, audio chroma vector is a 12-dimensional representation of the tonal content of an audio signal derived by combining bands belonging to twelve pitch classes. For the needs of this paper, our chroma feature vectors are obtained by using the toolbox provided in [14]. At the nth time-frame, the sub-fingerprint $\mathbf{F}_i(n)$ is 12×1 vector normalized so that $\|\mathbf{F}_i(n)\|_2 = 1$. We then proceed by considering $R_{\mathbf{F}_i,\mathbf{F}_j}(\tau) = c_{ij}(\tau)$, where $c_{ij}(\tau)$ is the number of the non-zero elements along a diagonal of a matrix derived by following the procedure described in section 5.2 of the work in [7]. Depending on the maximum diagonal score, pair ij is assigned a positive or negative match as explained at the beginning of section 2.

2.3. Wang's method

The fingerprinting method of Wang [1] was among the first ones to be used in the context of crowdsourced content management [5, 6, 8, 11]. The method operates by identifying local peaks in the TF domain called "landmarks". The combinatorial pairing of the landmarks into "hashes" significantly increases the robustness to noise and other types of signal degradation [1]. For implementing this method, we use the Matlab toolbox provided in [15] by considering $R_{\mathbf{F}_i,\mathbf{F}_j}(\tau)$ to be equal to the number of hashes matched as a function of the time-frame offset τ. In fact, the toolbox directly returns to us an integer w_{ij} which represents the number of maximum hashes matched, as well as the time-frame offset $\hat{\tau}_{ij}$ where this occurs.

2.4. Proposed approach

The feature vector that we propose takes into account energy variations in the frequency region from 0.5 to 1 kHz. We have observed that this frequency region represents the most energetic part of the spectrum for a chanting crowd event, a fact that is also demonstrated in the spectrogram of the sound signal in Fig. 1. For the fingerprint extraction process, we

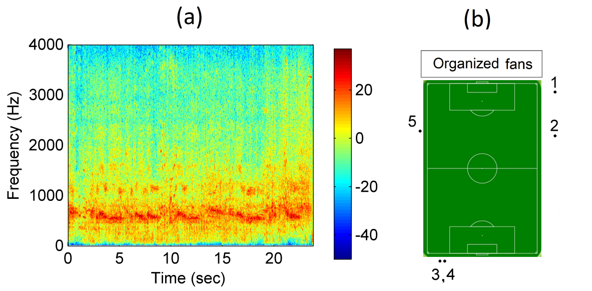

Fig. 1. Spectrogram of a chanting crowd event, as received at one of the smart-phone devices, in (a), and sketch of the stadium with corresponding recording locations denoted with black dots in (b).

need first define the auxiliary signal

$$P_i(n) = \sum_{m=n-L}^{n+L} \sum_{k=k_{LB}}^{k_{UB}} |X_i(m,k)|, \tag{4}$$

where $X_i(m,k)$ is the STFT coefficient associated with the mth time-frame and kth frequency bin of the audio signal at recording i, k_{LB} and k_{UB} are the frequency indexes corresponding to 500 and 1000 Hz respectively and L is a positive integer used for time averaging. The sub-fingerprint at time n is a scalar defined as

$$F_i(n) = \begin{cases} \text{1 if} & P_i(n) \geq P_i(n-1), \\ \text{-1 if} & P_i(n) < P_i(n-1). \end{cases} \tag{5}$$

Following the steps described in the beginning of section 2, a decision regarding a positive or negative match, as well as the time-frame offset required for synchronizing files i and j is taken upon the values of the generalized cross-correlation defined as

$$R_{\mathbf{F}_i,\mathbf{F}_j}(\tau) = \sum_{n=-\infty}^{\infty} F_i(n)F_j(n+\tau). \tag{6}$$

3. EXPERIMENTAL VALIDATION

While our scope is to present a fingerprinting process which is suitable for the organization of UGV acquired in a large scale sport event, we also present results for the case of a musical concert. The case of a musical concert has been investigated extensively in several studies in the past, but we believe that it adds further value in the context of this paper, by highlighting the inherent differences between those two types of acoustic events and the way that this is reflected upon the audio matching metrics provided by each audio fingerprinting technique.

The data required for the experiment was acquired as follows. For the athletic event, 5 participants contributed video recordings using their smart-phones. Their locations, which were fixed through the entire duration of the game, are illustrated with black dots in Fig. 1(b). Throughout almost the entire duration of the game, a big part of the crowd was chanting, mainly following the organized fans of the hosting team, distributed at the north side of the stadium, as shown in Fig. 1(b). On the other hand, the data in the music event was extracted from 7 different recording devices, at locations which were static during each song but which varied from one song to the other. The public address system located at the left and right side of the stage, represented the most dominant acoustic source during the concert.

More details with respect to the recording process in each event can be found in [16] while access to the datasets is provided by the link in [17]. We briefly note here that in total, 41 and 50 audio recordings were used for the case of the concert and the football match respectively, corresponding to 77 and 92 cases of positive match and 743 and 1133 cases of negative match. Finally, the average overlap duration between positive match pairs in both event was equal to 2 minutes approximately.

For all fingerprinting methods we used a sampling rate of 8 kHz with an analysis frame of 512 samples length. A hop size of 128 samples was used for all approaches except from Wang's algorithm which was implemented with a hop size of 256 samples. Finally, the value of L in Eq. (4) was set to 6. Using the audio, the video and the metadata which was available for each file, it was easy to identify whether a pair of recordings overlapped in time or not and this was used as the groundtruth for evaluating the matching performance.

For illustrating the matching performance in relation to each fingerprinting algorithm we decided to create two separate histograms with the match strength values, one for the cases of negative and one for the cases of positive match pairs. With respect to the concert, these histograms are shown for the three state-of-the-art techniques and for the proposed technique in Fig. 2. For better clarity, the plot is shown with a non-linear horizontal axis. It can be seen that the blue cardinality values corresponding to pairs of positive match are generally well separated from the negative match cardinality values, plotted in red. Also, by simple visual inspection, PRH and Wang's method in Fig. 2(a) and 2(b) provide better discrimination in comparison to chroma vector, where the overlap between red and blue bars is evidently greater at the intermediate match strength values in Fig. 2(c). The discrimination between the positive and negative match pairs is also very good with the proposed technique in Fig. 2(d). In terms of the $\{\text{PMS}, \text{NMS}\}$ metric, PRH and the proposed technique achieve the best scores of $\{98.7\%, 100\%\}$ for an empirically defined threshold of $\theta = 7 \cdot 10^3$ and $\theta = 7.5 \cdot 10^3$ respectively, both making one only false negative decision. The respective values for Wang's method and chroma vector are $93.5\%, 99.9\%\}$ for $\theta = 10$ and $\{83.1\%, 98.1\%\}$ for $\theta = 9$ respectively.

Contrary to the case of the concert, the distribution of the

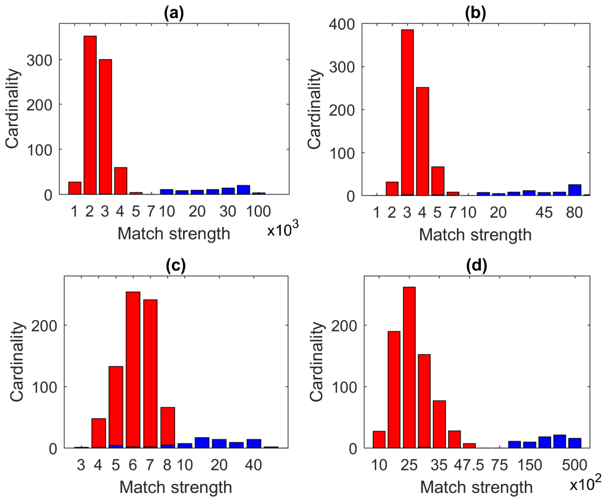

Fig. 2. Histogram with the distribution of the match strength, for the positive (blue) and negative (red) match pairs, in the case of the concert, for PRH in (a), Wang's method in (b), chroma vector in (c) and proposed fingerprinting technique in (d).

match strengths derived from the three state-of-the-art techniques in the case of the football match are very little informative about the actual match class, as can be seen in Fig. 3(a)-(c). For the majority of the positive matches, it is difficult to discriminate them from the negative match pairs, as their strengths are in the same range of values. Interestingly, some overlapping pairs appearing to exceed the match strength values of 15 and 7, in Fig. 3(a) and (b) respectively, associate to pairs from recording locations 3 and 4, which are at a very small distance as shown in Fig. 1(b). To our opinion, this is evidence that the state-of-the-art fingerprinting techniques detect a match because of the common acoustic "foreground", rather than because of the common background. To give an impression about the decline of the performance in the case of the football match, we note that, for an NMS of 97%, the PMS values are now equal to 19.6%, 19.6% and 17.1% for PRH, Wang's method and chroma vector respectively.

Finally, from Fig. 3(d) it becomes apparent that the proposed technique provides a substantially better audio matching metric, by clearly discriminating negative from positive matches. For a threshold value of $\theta = 900$, a PMS of 98.9 % with an NMS of 99.4% is achieved. This proves that the acoustic feature designed in section 2.4 correlates better to the common acoustic background and at the same time, manages to absorbs the variance characterizing channel pairs as far as 100 meters apart.

It is worth noting that the proposed approach not only solved the audio matching problem for the particular football match collection, but also provided with correct time-frame offsets $\hat{\tau}_{ij}$. In fact, by listening to the overlapping region of the time-aligned audio streams, it was confirmed that the estimated time-frame offsets were correct in all cases of positive match. Certainly, a decline in the performance should be expected for shorter overlap durations than the ones obtained in these specific experiments, but such an investigation is outside of the scope of this paper.

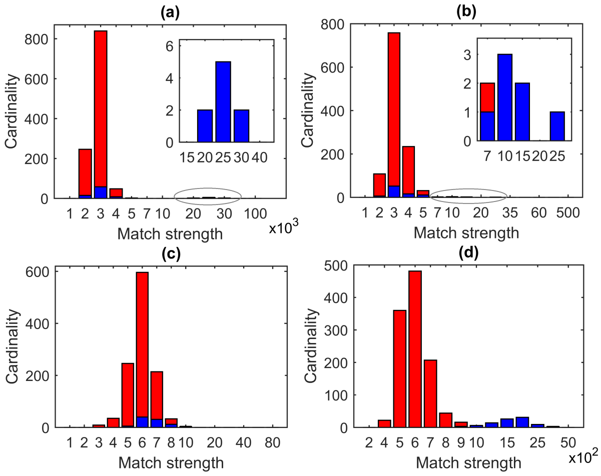

Fig. 3. Histogram with the distribution of the match strength, for the positive (blue) and negative (red) match pairs, in the case of the athletic event, for PRH in (a), Wang's method in (b), chroma vector in (c) and proposed method in (d). For better illustration, parts of the histogram are shown with a zoomed scale in (a) and (b).

4. CONCLUSION

Using an acoustic feature designed to correlate with the chanting of the crowd, this paper demonstrated the ability to automatically match and synchronize crowdsourced content acquired during a large scale athletic event. On the other hand, three of the most well known fingerprinting techniques failed in providing reliable matching decisions for organizing the same UGV collection. To our opinion, there are three basic differences in the acoustic conditions associated to the case of a concert and an athletic event; First, there is an enormous number of uncorrelated sound sources (the spectators) rather than a few number of dominant sound sources. Second, the sound sources are distributed over a large spatial region rather than being concentrated at one or a few specific locations. Finally, music has a higher rate of change, and goes through more drastic frame-to-frame changes than the sound of a crowd chanting in the football stadium. As a general conclusion, one should not expect a single fingerprinting technique to cope well with all types of crowdsourced content, as different acoustic features are more suitable than others, depending on the nature of the public event.

5. ACKNOWLEDGEMENT

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687605, Project COGNITUS.

6. REFERENCES

[1] A. Wang, "An industrial-strength audio search algorithm," in *Proceedings of the 4th International Conference on Music Information Retrieval*, 2003.

[2] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *International Society for Music Information Retrieval (ISMIR)*, 2002, pp. 107–115.

[3] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations.," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[4] P. Shrestha, M. Barbieri, and H. Weda, "Synchronization of multi-camera video recordings based on audio," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 545–548.

[5] L. Kennedy and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," in *Proceedings of the 18th international conference on World Wide Web*, 2009, pp. 311–320.

[6] C. Cotton and D. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 2386–2389.

[7] S. Bano and A. Cavallaro, "Discovery and organization of multi-camera user-generated videos of the same event," *Journal of Information Sciences*, vol. 302, pp. 108–121, 2015.

[8] J. Bryan, P. Smaragdis, and J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 2389 – 2392.

[9] M. Kim and P. Smaragdis, "Collaborative audio enhancement using probabilistic latent component sharing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 896 – 900.

[10] M. Kim and P. Smaragdis, "Collaborative audio enhancement: Crowdsourced audio recording," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 41 – 45.

[11] T. Hon, L. Wang, J. Reiss, and A. Cavallaro, "Fine landmark-based synchronization of ad-hoc microphone arrays," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 1331 – 1335.

[12] M. Müller, F. Kurth, and Clausen M., "Audio matching via chroma-based statistical features," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 288 – 295.

[13] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. IV–1429 – IV–1432.

[14] "Chroma feature analysis and synthesis," http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/.

[15] "Robust landmark-based audio fingerprinting," http://labrosa.ee.columbia.edu/matlab/fingerprint/.

[16] N. Stefanakis, S. Chonianakis, and A. Mouchtaris, "Two open access datasets of user generated audio recordings," https://doi.org/10.5281/zenodo.167311.

[17] N. Stefanakis, S. Chonianakis, and A. Mouchtaris, "Two datasets with user generated audio recordings," https://doi.org/10.5281/zenodo.164175.