The use of Human Power on Video Synchronization

Removed for Blind Review

ABSTRACT

User Generated Videos are contents created by heterogeneous users around an event. Each user films the event with his point of view, and according to his limitations. In this scenario, it is impossible to guarantee that all the videos will be stable, focused on a point of the event or other characteristics that turn the automatic video synchronization process possible. Focused on this scenario we propose the use of crowdsourcing techniques in video synchronization (CrowdSync). The crowd is not affected by heterogeneous videos as the automatic processes are, so it is possible to use them to process videos and find the synchronization points. In order to make this process possible, a structure is described that can manage both crowd and video synchronization: the Dynamic Alignment List (DAL). Therefore, we carried out three experiments to verify that the crowd can perform the proposed approach: the first experiment used a crowd simulator to verify the DAL capability of managing videos and contributions, generating cohesive video presentations; the second experiment used a crowd to synchronize videos performing small tasks; the third explored the use of the crowd to synchronize Live Stream Videos, through the development of the LiveSync tool.

CCS CONCEPTS

•**Information systems** → *Multimedia information systems;* •**Human-centered computing** → *Collaborative and social computing;*

KEYWORDS

Synchronization, Video, Crowdsourcing

ACM Reference format:
Removed for Blind Review. 2017. The use of Human Power on Video Synchronization. In *Proceedings of ACM Multimedia Systems, Taipei, Taiwan, June 2017 (MMSys'17),* 9 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Nowadays users don't need to rent fancy and expensive devices, neither depended on professionals to produce and share video, they can with their mobiles create User Generated Video (UGV). UGV is a kind of multimedia content created by heterogeneous users, shared online and without any explicit coordination mechanism. UGVs can be grouped around an event: a particular geographical space shared by a group of people at a particular period of time, such as protests, music festivals or sport games, resulting in awesome enahnced contents that can be reused in many ways. Moreover, each one of these videos reveals a unique point of view about what is happening, according to the userfis identity and beliefs (in terms of ideology, team and group identification etc.) as well as userfis context and preferences (in terms of positioning, device capabilities and limitations etc.).

In this scenario, it is impossible to ensure that all the UGV related to a same significant moment in a social event will be stable, have similar visual and aural quality or even if this moment was captured by the user. Then, automatic video synchronization techniques are not so effective. The synchronization of UVG about a particular topic or event can be announced as storytelling problem. Intuitively, this problem can also viewed as a synchronization problem, in which all of the related contents must be, firstly, positioned in a same global timeline and, in the sequence, arranged to produce a coherent narrative flowing.

Focused in this scenario, this paper introduces the idea of using crowdsourcing techniques for UGV synchronization, once the human processing is less affected by the heterogeneity of UGV content. Hence, we claim that the process of finding the synchronization points between UVG should involve crowd workers when automatic techniques are not effective. Also, the announced synchronization problem must consider many issues and in this paper, we focus on a particular issue: how to manage both crowd and video synchronization information and use them to rebuild the story of an event. In this sense, the paper introduces the Dynamic Alignment List (DAL), a data structure to assist the crowd contributions process as well as the generation of a coherent presentation of an event using UVG content.

We conducted three experiments to to investigate if the crowd would perform the proposed approach: the first used a crowd simulator to verify the DAL capability of managing videos and contributions, generating coherent presentations from user generated videos; the second experiment used a crowd to synchronize videos performing small tasks. The LiveSync Tool, developed to test if the crowd can be used to synchronize live streaming events, is detailed and shows how to use the structure to synchronize live strems. Last but not least, the focus of the paper is to show that the crowd can be used in scenarios where automatic techniques may find challenges, not to state that the use of the crowd is better than the automatic techniques.

2 CROWDSOURCING

Crowdsourcing is often a highly structured process from an organization, drawing on the creativity and intelligence of an online community in an open, but controlled, way [2]. Crowdsourcing systems can be used in a variety of situations such as: Knowledge Discovery and Management; Distributed Human Intelligence Tasking Organization; Broadcast Search; Peer-Vetted Creative Production [2]. The CrowdSync problem can be classified as a fiDistributed Human Intelligent Taskingfi problem, where an organization sends task related to information analysis. This kind of system deals with

MMSys'17, Taipei, Taiwan

DOI: 10.475/123_4

an information management problem, where the crowd will use information contained in videos to generate synchronization points. A second characteristic for our problem is that we consider that the information is already available and it is unnecessary to locate it. We don't approach the problem of collecting the videos, here we consider only synchronizing them.

Synchonizing videos isn't the only way to use the crowd with videos. Diverse works use the crowd with other objectives on videos. Two main problems may be highlighted here: annotation and quality evaluation, but diverse works can be addressed.

2.1 Annotation

The paper Video Summarization via Crowdsourcing [19] uses the crowd to identify the main occurrences in a video and thus generating a summary. In Efficiently Scaling up Crowdsourced Video Annotations [17] the crowd is used to annotate multiple videos. Through an interface provided by the authors, users can annotate videos using graphic tools, classifying objects in a video. Besides the annotation, which uses crowdsourcing, the paper presents a complementary approach that helps in annotating videos. Authors use annotations from the crowd as an input to an algorithm that automatically finalizes annotations from a video.

Generating Annotations for How-to Videos Using Crowdsourcing [10] also uses the crowd to generate annotations, however, in this case within a specific context. Authors divide the annotation task in three steps: identifying the time when important events occur; name each event; and identify frames indicating instants before and after the event, helping in the How-To.

Crowdsourcing event detection in YouTube video [15] focuses in generally identifying events in a YouTube video. It features processing in video playback time, to identify scene changes, and then allowing a viewer to annotate that part of the video. However, as stated by authors: fiRegarded in isolation, neither of our video event analysis steps is newfi. Its contributions are in: (i) scalability through crowdsourcing, (ii) the nature of real time processing in a HTML5 client, and (iii) the combination of annotations for three different types of events (visual event, occurrence event and interest-based event).

A web-based video annotation system for crowdsourcing surveillance videos [5] presents a platform for annotating surveillance videos. A supervisor selects and assigns tasks to users. The paper, despite using the term crowdsourcing , uses the concept of outsourcing, once results from the crowd are not evaluated, neither combined to generate the result. Each individual has to watch the entire surveillance video and annotate it, falling back to the large tasks problem.

Tagging human activities in video by crowdsourcing [11] uses the crowd to annotate scenes in a video, where each participant must annotate its start, end and details of its content to generate an annotation.

2.2 Quality

In Quantification of YouTube QoE via Crowdsourcing [6] the crowd is used when evaluating YouTube's quality of services (QoS). Each crowd individual watches videos from his home and rates his experience. An important point of the paper is the crowd filtering: in multiple tests executions, nearly 80% of them were removed from results as they were considered to fail the evaluation. For the elimination, these techniques were applied: Golden Standard Data; Consistency Tests; Content Quests; Mixed Answers; and Application Usage Monitoring. Video quality evaluation in the cloud [9] also presents a quality evaluation, but this time an evaluation from the actual video.

A subjective evaluation using crowdsourcing of Adaptive Media Playout utilizing audio-visual content features [12] presents another paper that uses the crowd to evaluate video quality. In this case, issues concerning adaptive video are evaluated. It is asked to participants to evaluate video quality, without knowing if video quality was modified or it remains the same.

2.3 Others

Crowdsourced Automatic Zoom and Scroll for Video Retargeting [3] uses the crowd to identify focusing regions of a video. Viewers select a zoom area that focuses important content from the video, so when the video is watched in low resolution screens, only most important content is presented to users.

Introducing game elements in crowdsourced video captioning by non-experts, [8] presents the use of game elements (gamification) in a crowdsourcing platform. In this case, the problem that needs to be solved is generating captions for multiple videos. The most significant about this paper is the use of video segments. This allows users to perform small tasks to achieve a greater goal. Gamification aspects are important as an incentive to participation, instead of monetizing tasks.

3 VIDEO SYNCHRONIZATION

Besides studying the relation between crowds and videos segments, it is important to consider other video synchronization techniques, and that is what we now describe now in this section.

The audio analyses can be used to synchronize the video segments. Su *et al.* [16] presents video synchronization using this approach. Fingerprints are generated for each audio, and a comparison between videos is performed. Bano *et al.* [1] also use audio as synchronization track using the chroma analysis from audio to group and synchronize audio from a same event.

There are also the approaches where video analyses is used instead of audio. Wang *et al.* [18], synchronizes videos in space and time, allowing teh naivigation between videos by resemblance and time. The synchronization works for multiple videos and different cameras. Other important works is described by Schweiger *et al.* [13] that a research for related papers in the area. It presents important results in automatic video synchronization, such as a technique that analyses differences between frames to find synchronization points. Furthermore, it presents the main challenges for automatic synchronization algorithms: wide baselines, camera motion, camera shaking, dynamic backgrounds and occlusions.

The use of human perception however, is not impaired by these. The human processing can overcame all these challenges and with or without the sound information. This is the principle that guided us to the proposed CrowdSync. Here we don't claim that the CrowdSync overcomes the automatic techniques, but with the

crowd we find less limitations on what videos we can synchronize.

4 CROWDSYNC

A traditional video presentation involves a single User Device that is able to decode and present this single content (the Main Content) originated from a unique source. In a non-traditional scenario [7], the presentation environment is composed of multiple user devices (TV, smart phones, tablets, etc.) able to present multiple contents delivered for multiple sources. One such scenario is an user that access a web video and is able to access other correlated videos with different angles, audio and complementary information.

In this situation, the user accesses a mashup of digital contents that may have no explicit synchronization defined to orchestrate the presentation. Mashups are applications generated by combining content, presentation or other applications functionalities from disparate sources. They aim to combine these sources to create useful new applications or services to users [21]. This combination of services and contents however brings the following issue: how to synchronize these multiple contents for each user in its environment, since the contents are transmitted through different channels and from different sources that are not explicitly synchronized among them? To tackle this issue we use the crowd as part of our solution. They act as couplers, in other words, they are responsible for finding the synchronization points among related videos, allowing their synchronous presentation in a mash-up video application.

The goal of video synchronization is to align a set of videos α in a common temporal line [14]. For this purpose, consider $\alpha 1$ and $\alpha 2$ as two continuous videos. They are considered to be synchronized when, in a given time T_k at the kn^{th} time instant of $\alpha 1$ and T_M at the mn^{th} time instant of $\alpha 2$, they both correspond to the same instant in global time when they were captured, which is an instance of continuous space-time. If they are not synchronized, there is a time offset (Δ) that added to the presentation of $\alpha 1$ or $\alpha 2$ will make them synchronous. The time offset between two videos V1 and V2 can be defined as $\Delta_{V1,V2} = b.V1 - b.V2$, where "b.V" is the starting time of video V in reference to a timeline of the related videos, and Δ is the time offset between them in this timeline. Finding this Δ is the task attributed to the crowd. They are responsible for analysing the videos, finding the correlated ones and setting the Δ that makes them synchronous.

Three different Crowd Synchronization scenarios are presented next: Chunk Synchronization, Frame Synchronization and Live Synchronization.

4.1 Chunk Synchronization

Following the crowdsourcing approach, we can't let each crowdworker analyse all the entire videos to find all synchronization points. This would require too much effort of each crowdworker. For the current version of our CrowdSync system, we split the videos in small chunks of 5s each. This way we make each task a lot easier to each crowd member, because for each task he needs only to compare if two chunks overlaps and if they do, what is the Δ that makes them synchronous.

Figure 1 shows this synchronization method. First each video A e B is mapped in chunks of 5s. A pair of chunks is sent to a crowdworker that evaluates if there is synchronization and what is the Δ if there is any. In the example (Figure 1), if the chunks $[C_1A,C_1B]$ have no synchronization, we compare the next possibility $[C_1A,C_2B]$ and so on until we find it or we compare all chunks. In the example the pair $[C_1A,C_3B]$ contains a synchronization point. The crowdworker identifies it and discovers the time offset between them (Δ). Using the value of Δ and knowing which chunks where the ones where the synchronization point was found, we can synchronize the full videos [A,B]. In the example, the final difference between the two videos [A,B] is Δ plus two times the chunk size, because Δ was found in the relation between the first chunk of A and the third of B, a difference of two chunks.

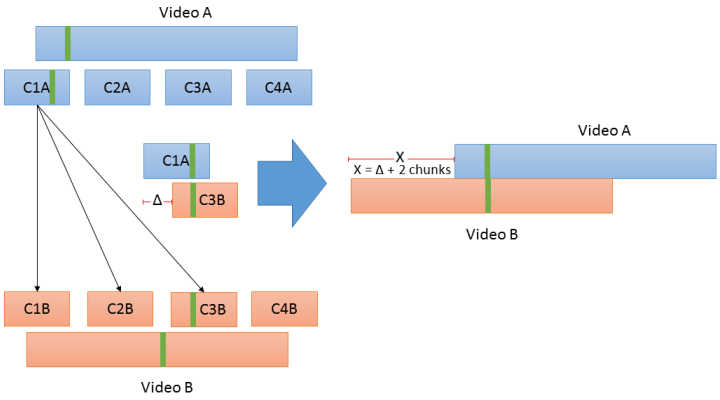

Figure 1: Chunk Synchronization Method

When two chunks associated with two videos are synchronized, all the contents of these two videos will also be synchronized. This happens because as described in the introduction, we consider the videos as continuous. And if comparing all chunks from both videos, we can't find any synchronization point, those two probably have no synchronization point. We say probably, because there is a possibility that the crowd fails in its tasks. To reduce this possibility, there are measures that can be taken such as assessing the crowd for better contributions. We however do not discuss this issue in this paper, as our scope focus on the synchronization process only. Another important consideration here is that this approach can fall in the worst case when there is no synchronization point, and all chunks will be compared, and will fail. On the other hand, if there is an synchronization point, we can find it in the first comparison.

4.2 Frame Synchronization

The Frame synchronization is an enhancement proposal on top of the chunk model. Here, instead of using video chunks, each crowdworker receives the key-frames sequences from two videos that must be synchronized. This way, each worker has access to the full video at once, enhancing the possibilities of finding a synchronization point. However, as now the worker interacts with frames, if a pair of frames is identified as a possible synchronization point, its precision is reduced, making necessary a second step on the task: watch the videos based on the information provided by the

two frames (each frame has a timestamp from when it which part of the video it was removed, making possible to calculate Δ such as in the chunk based one).

Figure 2 shows the interface developed to this approach: on the top the user selects the probably aligned frames (or close ones) and plays the videos. If it is correct, he confirms the alignment, else he tries again or discard the synchronization.

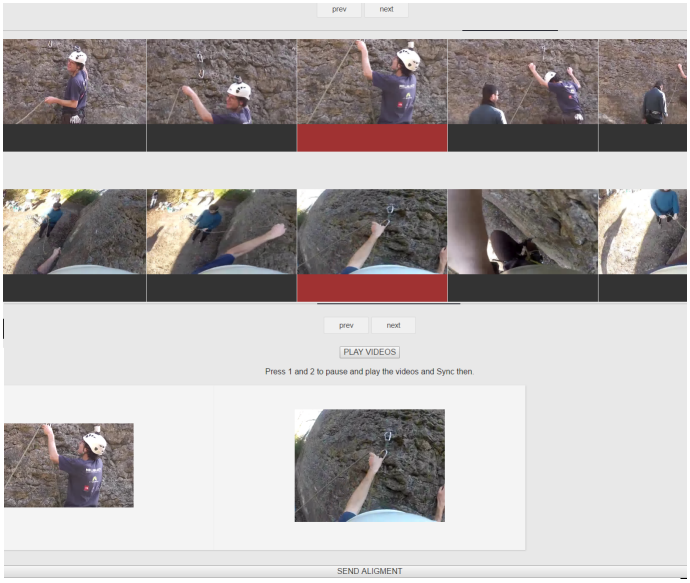

Figure 2: Frame Synchronization Method

4.3 LiveSync

Live synchronization includes the scenario where a viewer has access to an event that is live streamed by more than one Content Provider. These content providers are independent, so their videos do not have initial resources that allow their automatic synchronization to viewers, requiring a video analysis to generate synchronization points (Couplers).

As example for this scenario, we can take a public manifestation. In the event, multiple people can take their cell phones and start streaming the event. In their house, other people can watch the videos. However, the multiple videos from different sources will be asynchronous. We need a way to synchronize these UGVs. We use the crowd to achieve it. To this objective we group all videos in a MashUp application that connects to a Coupler Server that contains all synchronization data. Both synchronizing and playing the videos are made using this mashup, that can receive videos from multiple sources.

In a live presentation, it is assumed that content must be consumed right after its generation. It is of extreme importance that the synchronization method can be performed in playback time, to allow the integration of live content. However, not all viewers are required to be part of the crowd to achieve synchronization. If a synchronization made by a single member of the crowd is accepted as accurate, it can be transferred to the remaining viewers, this way each one will have his content locally synchronized.

One way of live synchronization can work is as follows: a person selects and synchronizes two videos with the help of a manipulation tool. This becomes a candidate synchronization point. Several people can do the same, and the results can be based on multiple synchronizations. Having these synchronization points defined, synchronization information can be sent to other viewers interested in watching those videos. As simple example, take two independent sources that are transmitting an event. A mash-up system allows the user to watch both videos at the same time is his device. However the videos are asynchronous and the user notes that. He then access the option to synchronize the videos. After he achieve a synchronous result, implicitly his contribution is sent to a server that will feed other users that choose to watch the same videos with the synchronization specification. If the user thinks the content is not synchronised yet, he can synchronize it himself and send another contribution. This tool was implemented and is presented on section 6.1.

5 DYNAMIC ALIGNMENT LIST

Besides the synchronization method, there is the need to manage all the generated tasks: which pair of chunks will be sent to each user? From which videos? Where are the contributions stored? How can the contributions be validated? How to know that all videos are synchronized? What are the time offsets among the videos? Do I need to compare all videos?

Trying to answer these questions, we developed the concept of a Dynamic Alignment List. It is responsible for managing all contributions, checking the convergence of the contributions, distributing the videos, inferring unknown values and storing Δ values for each pair of videos.

The DAL has two main features: (i) time offset and (ii) contribution managements. The time offset management deals with all videos relations manipulation while the contributions management takes care of the crowd management.

5.1 Structure

The DAL isn't just a list as the name may suggests, it is a data structure composed of vectors and lists that are used to assist the processing of contributions of crowd workers as well as to support the generation of presentations with the synchronized videos. It was designed in order to efficiently store all relevant contributions that represent a temporal relation between a pair of videos.

Its structure is based on a Upper Triangular Matrix, although, the DAL is constructed on demand, saving space because it doesn't have any empty slot on it, turning it into a dynamic structure. In Figure 3 it is shown an instance of a DAL with five videos and its relations.

The DAL starts as an array that contains in its positions the videos that are going to be related (blue squares). Each video is an object that contains a reference to the media that it represents (URI) or other DAL (creating a hierarchy), a reference to the array of relations with the other videos and an attribute with the duration (σ) of the video.

These relations are cells into an array. Each position of this array is a different relation that represents the time offset between those two videos. Each position of the array stores the IDs for the pairs of

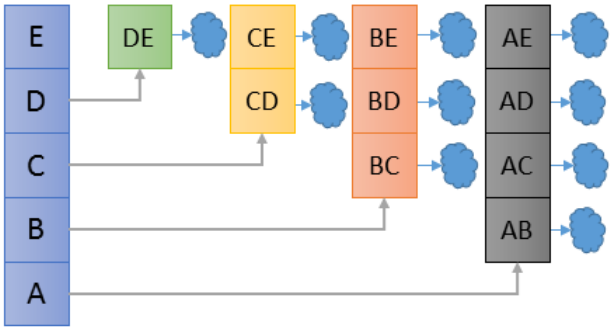

Figure 3: DAL Structure

videos that are being related, an attribute that contains the current Δ (time offset), an attribute that contains the degree of confidence in that relation and a list of contributions on that relation (each contribution on which is the Δ value). The degree of confidence is how precise the current Δ is believed to be true. Note that we don't need to store all direct relations (e.g. $\Delta_{A,H}$ and $\Delta_{H,A}$), we can store one direct relation and its complementary can be easily calculated ($\Delta_{A,H} = -\Delta_{H,A}$).

A contribution (in the blue cloud) in turn has the Δ proposed by each user for that relation and a reference for that user (the user profile may be used to rise the degree of confidence of a relation).

5.2 Time Offset Management

The first feature (and original purpose) of the DAL is the Time Offset Management. It is responsible for dealing with all aspects of the time offset relation among videos. These aspects are:

Δ **Storage and Retrieval:** a characteristic of the DAL is to allow the addition, update and retrieval of any Time Offset for any pair of videos. This gave us the starting point of using a Matrix like structure that permits fast access to any cell of the matrix from a pair of coordinates. This way knowing the two videos which we desire to know or to update the Δ, we can quickly obtain this information. However we don't use a pure matrix for two main reasons: information redundancy, we don't need to store the AB and BA relations, as one is the opposite of the other ($\Delta_{i,j} = -\Delta_{j,i}$, eq. 1); and a matrix doesn't allow the storage of multiple contributions in the same cell (requisite for the crowdsourcing);

$$\begin{aligned} \Delta_{i,j} &= b.i - b.j \\ \Delta_{j,i} &= b.j - b.i \\ -\Delta_{j,i} &= b.i + b.j \\ \Delta_{i,j} &= -\Delta_{j,i} \end{aligned} \tag{1}$$

Δ **Inference:** known relations in the DAL allows developing inference methods that use transitivity through pairs of aligned videos to calculate relations between videos,in which the offset is still unknown. If we know the relation AB and AC from the structure we can infer BC.

$$\begin{aligned} \Delta_{B,C} &= b.B - b.C \\ \Delta_{B,A} = b.B - b.A &\rightarrow b.B = \Delta_{B,A} + b.A \\ \Delta_{A,C} = b.A - b.C &\rightarrow b.A = \Delta_{A,C} + b.C \\ \Delta_{B,C} = \Delta_{B,A} + b.A - (b.A - \Delta_{A,C}) &\rightarrow \\ \Delta_{B,C} &= \Delta_{B,A} + \Delta_{A,C} \end{aligned} \tag{2}$$

Presentation Generation: With the known and inferred time offsets, it is possible to create a presentation of the event that correlates all videos with temporal relations. This presentation can be the maximal one, where the event is presented using the first video until the end of the last one, also when videos with overlapping exists, the option to change camera is made available to users.

5.3 Contribution Management

The Contribution Management is a requisite in order to adequate the DAL for a crowdsourced scenario. It is necessary to receive, distribute and process the crowd contributions in order to find the correct Time Offsets. The aspects involved in the Contribution Management are:

Convergence: one of the principles of crowdsourcing is collecting the contribution of multiple members and based on these contributions finding the solution to a problem. In our case, the crowd watches the videos fragments and find synchronization points. The convergence is responsible for getting all theses contribution, and merging them in actual results.

The Convergence Level detects the tendency of the Crowd about the Relation of a pair of videos. This tendency is an indicative of the agreement on a Delta between the videos. According to the scenario, a Convergence Threshold that is used to determine if a Relation is converged, is defined. Each Relation have an attribute that stores its current convergence level. If this level reaches a Convergence Threshold the method returns True, otherwise it returns False.

Initially all Relations are created with Convergence Level 0, and consecutive similar contributions increase this value. If a tendency change in the contributions is detected, the Convergence Level is reset.

Video Selection: the convergence deals with the problem of receiving and processing the crowd contributions. However, we must correctly choose the videos to be evaluated by the crowd that results in a contribution. The selection of these videos is part of the contribution management.

This method intent is to select which pair of videos should have priority to receive contributions. It consists of a sequence of two other methods: (i) Choose the next Video; (ii) Choose the next Relation. When this method returns NULL, it means that the DAL has been converged.

In order to increase possible inferences over the contributions, the method try to spread contributions over the timeline, proceeding a random selection among the Videos that were not converged nor marked as a Impossible relation. Once a Video is chosen the method select from its Relations array the Relation closer to converge.

6 EXPERIMENTS

Aiming to verify the CrowdSync method and the data structure generated to implement the method, two different experiments were executed:

LiveSync Tool: the LiveSync Tool implements all requisites for live synchronizing live UGV streams. Its objective is to allow the creation of mash applications of related video streams with client based solutions.

Crowd Simulated DAL: the second experiment uses a simulated crowd to verify the DAL. The objective here is to verify if the DAL correctly stores, infers, converges and selects videos and relations. We simulate the crowd so we can make a controlled analysis in a set of videos with different crowd profiles.

UGV Dataset Synchronization: in the third experiment, real users (crowd workers) synchronize a user generated video dataset using a developed platform with all characteristics described in this paper. Since the simulated experiment allows only to verify the correct functioning of the algorithms and structure, the focus here is to verify if humans using their perception can find the synchronization points in all related videos.

6.1 LiveSync Tool

The main functionalities provided by our tool are:

Synchronized Live Video Player: The tool permits users to watch multiple videos synchronized. He selects from a list of sources the videos he wishes to watch and them they are synchronized using information provided by other users. If a pair of videos does not have any information about their synchronization, users are invited to contribute and synchronize the videos.

Synchronization Infering: In some cases that there is no direct information about the synchronization of two videos, the tool is able to infer the synchronization about them, based on the contributions of other videos. We use the transitive attribute of video synchronizations, where if we know AB and BC synchronization info (couplers), we can infer AC. To infer this value we travel trough the DAL, finding the unknown relations (for example, CE), and try to find a path of known relations where we can infer CE. Taking the DAL in Figure 1, we can infer CE if we know: AC and AE. This a two steps route, but we try all possible routes when inferring, in a way that we fill as much relations as possible.

Video Aggregation: Although the focus of the LiveSync tool is on synchronization, we allow users to add new stream sources to the application. He only needs to set the video source, and the video will be added to the DAL and list of videos. However, we don't do any filtering about the added video, this means that the user can add any video to the application, even ones that contains none relation with the other videos. In future versions we plan to add options where other users can mark the video as not related, and then remove them.

Multiple Video Platform Support: One keypoint of our tool is the use of other platforms as video sources. The videos that we play to users and that are synchronized, are not provided by us, but by other live video stream platforms. To be compatible with our tool, two requisites are required:

(1) Remote Player: we need that the platform allows embeddable players on third pages, allowing us to control the player with its basic functionalities such as: play, pause and stop;
(2) Uptime Support: a second and fundamental requisite is an API that allows us to retrieve the video Uptime. Video uptime is the time since the beginning of the video that is presented on the video player. This is fundamental to create and replicate the couplers generated in synchronization process.

Serverless Architecture: Serverless architectures refer to applications that significantly depend on third-party services and putting much of the application behavior and logic on the front end. Such architectures remove the need for the traditional server system sitting behind an application.

Multiplatform: LiveSync is a Web Based application designed and developed in compatibility with HTML5 standard to its front-end (MashUp Player) component. It allows our application to run on multiple browsers, operational systems and devices.

Active X Passive Contributions: Two branches of the LiveSync are currently on our repositories. They differ only in one aspect: who defines what videos are to be synchronized: the crowd or the application? The active version allow users to navigate freely through the videos, synchronizing them when they wish to. The focus of this branch is to allow users to contribute if they want to. On the other hand, in the Passive branch the server gives the crowd exactly what video they will synchronize. The focus here is to rapidly synchronize all videos, so the focus isnfit to make users watch the videos, but force them to synchronize all the base for other porpoises. The active branch is the focus here, but can easily converted to the passive one.

6.1.1 Architecture. The LiveSync tool has three main components (Figure 4): the Content Providers (Video Sources), the Coupler and the Mashup Player.

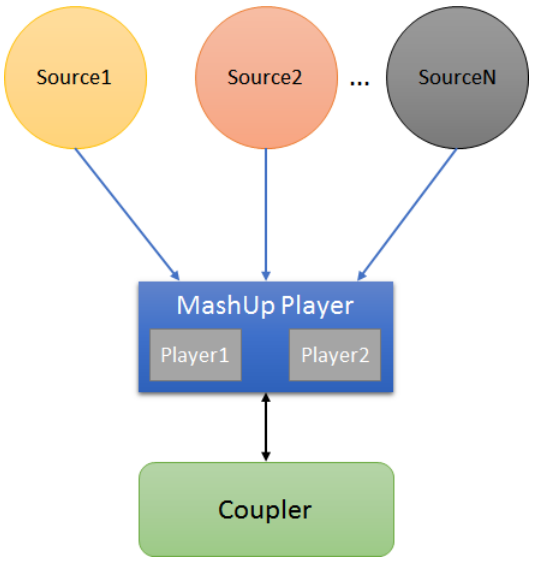

Figure 4: LiveSync Model

Content Providers

Content Providers are third-parties videos streamers platforms There are multiple Live Video Stream applications in the market, such as YouTube Live, LiveStream, TwitCast, Twitch and Ustream One of our objective was to allow the use of different platforms as video sources, so we maximize the number of videos for an event and allow the use of already in market platforms. A Content Provider needs two requisites to be compatible with LiveSync: a Remote Player and Uptime Support. As each stream platform uses their own protocols, we opt to use their embeddable players into a MashUp application. These players must allow us to play, pause and stop the video stream. The second requisite, Uptime Support, is necessary to find the couplers among the videos. Uptime is the time passed since the beginning of the live stream until the video part being presented in the player at the moment of the call.

Coupler

The Coupler is responsible for storage, distribution and calculation of synchronization points among video streams from the Content Providers.

A coupler is composed of a DAL instance and Log files. This goes in direction of the Serverless Architecture. We wanted an architecture that needed low resources (another justification for using third party stream services) and easy deployment. All that is necessary do execute the coupler is a NODE.JS (https://nodejs.org/en/) server instance. This is possible because the Coupler is fully developed in JavaScript and compatible with the HTML5 standards. To deploy the Coupler, we use a Backend as a Service or fiBaaSfi platform, more specifically we use the Heroku (www.heroku.com) one, that permits free use of NODE.JS instances.

It stores synchronization information only during the duration of the event, so its stance is finished with the end of the videos and all data is lost. In the current scope, the sync info is only necessary during the event, after it, there is no need to store the information. For reasons of testing and using the filmed videos from YouTube we create log files that contains all contributions made by the crowd. If it is important to maintain all contributions and data for post analyses and further use, unstable version of the LiveSync is being configured to use a fully transactional database. We use a fully transactional database because we want to maintain track of all contributions made by the crowd, an important aspect in crowdsourcing and that is also supported by the DAL.

MashUp Player

Mashups are applications generated by combining content, presentation or other applications functionalities from disparate sources. They aim to combine these sources to create useful new applications or services (the offer and consumption of data between two devices) to users. In LiveSync we combine videos coming from different sources and platforms with the synchronization information from the coupler to reproduce a synchronous presentation of these videos.

The MashUp Player (Figure 5) is responsible for both presenting video synchronously and collecting the synchronization. Figure 5 represents the interface of the MashUp during a test: two cameras live streaming (content providers) a simulated television event to our mash-up application.

On the top we have all information necessary to the user. He can aggregate new videos using the ADD NEW VIDEO options,

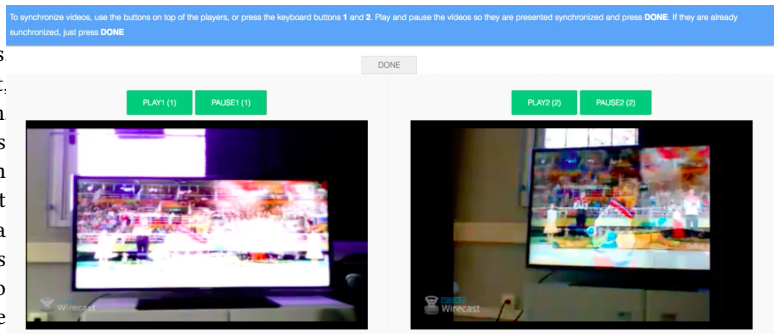

Figure 5: Live Streams from Olympic Games Synchronized through two different cameras

SYNCHRONIZE the videos if he thinks the videos are not synchronized or he can just select the videos he want to watch. When just playing two selected videos from the videos list, each video player creates an instance for the player that is compatible with that source (YouTube orWeb- Socket). It is invisible to the user where the video is coming from.

When the user adds a video, an input text is shown to him, and he can add the video URI (WebSocket) or video ID YouTube). The page reloads and the new video is listed in the video list for everyone that connects to the application. When the video is added by the user, an message is sent to the Coupler, containing the action to add a new asset to the DAL, and the specification of it, such as label and URI.

The last functionality of the MashUp is to synchronize the videos. When the user clicks on SYNCHRONIZE, a new mode of the application is revelled showing the synchronization tools. We use a Play fin Pause approach to synchronize the videos. After the user thinks the videos are synchronized, clicking the DONE button, his contribution is sent to the Coupler and stored in the DAL for further processing of the relation.

6.2 Crowd Simulated DAL

Verifying the DAL only with a real crowd in a real scenario could result in biased results, because of the human factor. We can't fully control the crowd in this scenario, so our test analyses could be compromised. Then to test the DAL and its features we decided to simulate the crowd contributions.

The first step here was to find the behaviour of the common crowd. Yu *et al.*[20] classify the crowd in four categories: Hon workers: honest worker agents who return high quality HIT (human intelligence task) results randomly 90% of the time; MH workers: moderately honest worker agents who return high quality HIT results randomly 70% of the time; MM workers: moderately malicious worker agents who return high quality HIT results randomly 30% of the time; Mal workers: malicious worker agents who return high quality HIT results randomly 10% of the time. In our scenario we consider a high quality HIT as a positive synchronization point identification.

The second step consists of defining a dataset. As no human will analyse the video, only the duration and times of the video matter to our problem. We created a simulated 600s-length event and built

Table 1: Median for 80 Videos, after 30 rolls of simulations for each Trustworthy degree

Reliability	Contributions	Direct	Inference	Errors
100%	2786	77	2628	0%
90%	3018	77	2602	0%
80%	3280	77	2578	0%
70%	3642	77	2559	0%
60%	3956	77	2538	2%
50%	4328	77	2525	2%
40%	4945	77	2507,5	2%
30%	5763	77	2471,5	5%
20%	6007	86	2412	9%
10%	7338	85	2357	10%
0%	8821	83	2258	16%

a dataset with 80 video segments with randomly generated videos varying from 5s to 300s over the event timeline, in a total of 9408s of video content to analyse. Based on these videos we built a Gold Standard DAL with 2766 Possible Relations and 394 Impossible Relations (gaps) to evaluate the results of our experiment.

The third step was to simulate the crowd synchronizing these videos. Yu *et al.*[20] presented a proportion of Hon, MH, MM and Mal worker agents varying the number of trustworthy workers from 10% to 100%. Trustworthy workers comprised of half of Hon and the other half by MH workers. The remaining non trustworthy workers are half MM and half Mal. Using these proportions high quality and low quality contribution were generated to the DAL until its convergence.

Table 1 presents, for each reliability degree of the crowd, the median of how many contributions were needed to converge the DAL (find all time offsets) after 30 roll of the simulation, considering that to agree on a value the Convergence Level required is 3 (at least three agreements are necessary to converge). The table also presents how many Relations were obtained by direct contribution of the crowd and how many were inferred by the DAL, and the number of errors (direct and inferred) comparing the results with the expected ground truth. It is noticeable that in the interval of 100% and 60% of Trustworthy workers the found relations is very similar (difference of 3%), but as the Trustworthy is reduced the more contributions are required (42% for the same interval).

6.3 UGV Dataset Synchronization

Instead of discarding the human factor, this experiment aims at testing if the human factor is able to find and synchronize an User Generated Video dataset. The Climbing video dataset [4] presents multiple videos filmed by a group during a climbing activity. The automatic technique presented with the dataset presents some limitations making this dataset ideal to test the crowd approach, so we selected part of the dataset (9 videos) to our experiment. The automatic method could identify 23% of all pairs of videos that temporally overlap, the pairwise alignment score (PAS).

To execute our experiment, we developed the technique described in section 4.1, where the crowd member analyse video chunks (5s) pair by pair, identifying if there is relation and the precision of the relation if one exists. This was the chosen solution because requires less effort from the crowdworkers, and more detailed tests are necessary to evaluate the Frame Synchronization one, as a more complex task may compromise the worker activity.

From the known ground-truth we analysed the values resulted from the crowd to evaluate the alignment within a tolerance of 0.5s [4]. The comparison with the ground truth resulted in 88% (PAS) of correct relations among the videos. The nine videos generated 278 chunks of video (a total of 1390 seconds to be synchronized) demanding a total of 1051 contributions. Most of these (99%) contributions were negative ones, in other words, the crowd worker could not identify synchronization point. This indicates that most of the crowd effort is being done in discarding synchronization point, and not really finding them.

Figure 6 shows an instant where overlapping occurs in videos recorded using 6 different cameras are synchronously presented after the crowd contributions. The other three ones are not shown because in this instant no overlapping was indicated by the crowd. It is possible to notice how heterogeneous are the camera shots.

Figure 6: Synchronized Video Matrix

7 FINAL REMARKS

Crowds can be used in the most diverse situations: from designing a product to digitizing a word, going through most diverse scenarios, as discovering the structure of a protein. Exploring this variety of uses of the crowd this paper presented the possibility of using crowds in the video synchronization process. Using the crowd allows to address challenges that automatic processing techniques struggle to solve, like: moving cameras, constantly changing backgrounds, disappearing objects and others. A human can handle this problems without affecting his perception about a video

Nevertheless, using crowdsourcing techniques also introduces new problems to the process: which pair of videos will be sent to each user? Where are the contributions stored? How can the contributions be validated? How to know that all videos are synchronized? What are the time offsets among the videos? Do I need to compare all videos? As solution to these problems the DAL was described. An structure that can manage temporal relations and crowd contributions.

The first experiment showed the simpler where the crowd can be used: the live one. It is said simpler because the crowd has a smaller search area to find the videos alignment point. With the LiveSync this scenario can be solved and multiple mashups applications developed. The Second experiment showed us that the more reliable the crowd is, the less contributions we need and that

100% to 60% may generate similar results. Also we don't need the crowd to find all values: most of the relations we can infer from others. The third experiment showed us that it is possible to use the power of the crowd to synchronize video datasets, in the a hard dataset that presented challenge even to automatic solutions. From the second experiment we also learned that most contributions are to find that two chunks don't have a synchronization point. This is important and is the key point to the next steps in the development of our research, as we are developing new interfaces that helps the crowd members in identifying the synchronization points in entire videos at once, not only chunks. This will lead us in reducing the number of required contributions and achieve more accurate results.

All code involved in our research is opensource and is available in [Removed for Blind Review].

REFERENCES

[1] Sophia Bano and Andrea Cavallaro. 2015. Discovery and organization of multi-camera user-generated videos of the same event. *Information Sciences* (2015).

[2] Daren C Brabham, Kurt M Ribisl, Thomas R Kirchner, and Jay M Bernhardt. 2014. Crowdsourcing applications for public health. *American journal of Preventive Medicine* (2014).

[3] Axel Carlier, Vincent Charvillat, Wei Tsang Ooi, Romulus Grigoras, and Geraldine Morin. 2010. Crowdsourced automatic zoom and scroll for video retargeting. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 201–210.

[4] Matthijs Douze, Jérôme Revaud, Jakob Verbeek, Hervé Jégou, and Cordelia Schmid. 2016. Circulant temporal encoding for video retrieval and temporal alignment. *International Journal of Computer Vision* (2016).

[5] Neeraj J Gadgil, Khalid Tahboub, David Kirsh, and Edward J Delp. 2014. A web-based video annotation system for crowdsourcing surveillance videos. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 90270A–90270A.

[6] Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. 2011. Quantification of YouTube QoE via crowdsourcing. In *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 494–499.

[7] Zixia Huang, Klara Nahrstedt, and Ralf Steinmetz. 2013. Evolution of temporal multimedia synchronization principles: A historical viewpoint. *ACM TOMM* (2013).

[8] Hernisa Kacorri, Kaoru Shinkawa, and Shin Saito. 2014. Introducing game elements in crowdsourced video captioning by non-experts. In *Proceedings of the 11th Web for All Conference*. ACM, 29.

[9] Christian Keimel, Julian Habigt, Clemens Horch, and Klaus Diepold. 2012. Video quality evaluation in the cloud. In *International Packet Video Workshop)*. IEEE.

[10] Phu Nguyen, Juho Kim, and Robert C Miller. 2013. Generating annotations for how-to videos using crowdsourcing. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 835–840.

[11] Long-Van Nguyen-Dinh, Cédric Waldburger, Daniel Roggen, and Gerhard Tröster. 2013. Tagging human activities in video by crowdsourcing. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 263–270.

[12] Benjamin Rainer and Christian Timmerer. 2014. A subjective evaluation using crowdsourcing of Adaptive Media Playout utilizing audio-visual content features. In *IEEE Network Operations and Management Symposium*. IEEE.

[13] Florian Schweiger et al. 2013. Fully automatic and frame-accurate video synchronization using bitrate sequences. *IEEE Transactions on Multimedia* (2013).

[14] Ricardo Segundo and Celso Santos. 2015. Remote Temporal Couplers for Multiple Content Synchronization. In *IEEE ICCIT*. IEEE.

[15] Thomas Steiner, Ruben Verborgh, Rik Van de Walle, Michael Hausenblas, and Joaquim Gabarró Vallés. 2011. Crowdsourcing event detection in YouTube video. In *10th International Semantic Web Conference (ISWC 2011); 1st Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*. 58–67.

[16] Kai Su, Mor Naaman, Avadhut Gurjar, Mohsin Patel, and Daniel PW Ellis. 2012. Making a scene: alignment of complete sets of clips based on pairwise audio match. In *ACM International Conference on Multimedia Retrieval*. ACM.

[17] Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* (2013).

[18] Oliver Wang et al. 2014. Videosnapping: Interactive synchronization of multiple videos. *ACM Transactions on Graphics* (2014).

[19] Shao-Yu Wu, Ruck Thawonmas, and Kuan-Ta Chen. 2011. Video summarization via crowdsourcing. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1531–1536.

[20] Han Yu, Zhiqi Shen, Chunyan Miao, and Bo An. 2012. Challenges and opportunities for trust management in crowdsourcing. In *IEEE/WIC/ACM WI-IAT*. IEEE Computer Society.

[21] Jin Yu et al. 2008. Understanding mashup development. *Internet Computing, IEEE* (2008).