Spatiotemporal Video Synchronisation by Visual Matching

Marcus Thaler, Werner Bailer, Georg Thallinger
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17
8010 Graz, Austria
{firstname.lastname}@joanneum.at

ABSTRACT
The media coverage of live events can be turned into a more immersive experience if content from multiple sources, e.g., professional and user generated content, are combined. We have implemented a visual matching approach to establish or improve temporal and visual synchronisation of such heterogeneous content. The approach is based on matching of SIFT descriptors and is implemented on the GPU. In order to visualise and explore the matching results, we have implemented a web-based viewer for the aligned content.

ACM Classification Keywords
H.5.1 Information Interfaces and Presentation: Multimedia Information Systems; I.2.10 Artificial Intelligence: Vision and Scene Understanding—*Video analysis*

Author Keywords
Visual matching, multi-view, user generated content

MOTIVATION
Many cultural and sports live events do not take place at only a single spot, but are spread out over different stages, halls, cities or even regions, with different actions happening in parallel in each of these places. Examples are music festivals with several stages or tents, city festivals, parades, marathons or bike races. Except for some few high-profile events, it is not possible to fully cover such events with professional capture equipment. In order to enable immersive coverage of this type of events, the ICoSOLE project[1] is developing technologies for live capture and streaming from professional and consumer devices, fusion of audio and video content from heterogeneous devices into a format agnostic representation, and methods for analysing and filtering streams based on quality and content properties.

Providing an immersive experience to the end user requires that the content is not only temporally synchronised, but also spatially aligned. This enables making appropriate transitions and supporting editorial staff in content selection by knowing which set of content showing a particular part of the scene is available. While we can obtain temporal synchronisation and precise location data from high-end equipment, this is a much more challenging problem for user generated content (UGC). We have developed a dedicated capture app that can take care of temporal synchronisation and captures a range of sensors, but this does not fully solve the problem. Absolute spatial localisation information from mobile devices may be unreliable indoors, and aggregated relative motion measurements may drift considerably over time. Some devices lack certain types of sensors, or users choose to deactivate them. In addition to mobile devices, there is consumer grade and semi-professional equipment such as DSLRs or action cameras, which provide good image quality, but lack in most cases the option for recording location data. And even if we have rather precise location information, we still need knowledge about orientation and zoom settings in order to know what is actually depicted.

In order to address this issue, we have implemented a visual matching approach to establish or improve temporal and visual synchronisation. We then describe a web-based visualisation of the matching results, which we have developed to validate and navigate the results.

APPROACH
For every 5^{th} frame of the videos we detect up to 5,000 key points and extract SIFT [3] descriptors for these key points. We use our GPU accelerated implementation of the SIFT (Scale Invariant Feature Transform) extraction pipeline [2] for this purpose. Then we compute pairwise similarities between each frame of the reference video and each frame of the UGC videos. Again, a GPU accelerated implementation of SIFT descriptor matching is used [1]. The key point matches between a pair of frames are validated by selecting the maximum number of descriptors supporting a consistent homography between the views. Videos with no or minimal overlap will not result in a significant number of matches.

In order to temporally align every video stream (typically the UGC streams) to the reference (typically professionally captured) stream, sequences of temporally adjacent frames with a high similarity are determined. Therefore, we build a similarity matrix of pairwise matching scores, and search for lines of high matching scores of a certain slope (determined by the ratio of the frame rates of the videos, i.e., diagonal in the case of identical frame rates). These lines may have gaps (if one or more frames in a sequence do not match well) and may have fuzzy start and end points. False positive matches between individual frames will result in scattered and isolated matches rather than sequences of matches and can thus be filtered. Often sequences are not unique for a certain period of time, e.g., when similar shots in a video exist. We then select the best match under the assumption that the incoming video

[1] `http://www.icosole.eu`

3rd International Workshop on Interactive Content Consumption at TVX15, June 3rd, 2015, Brussels, Belgium.

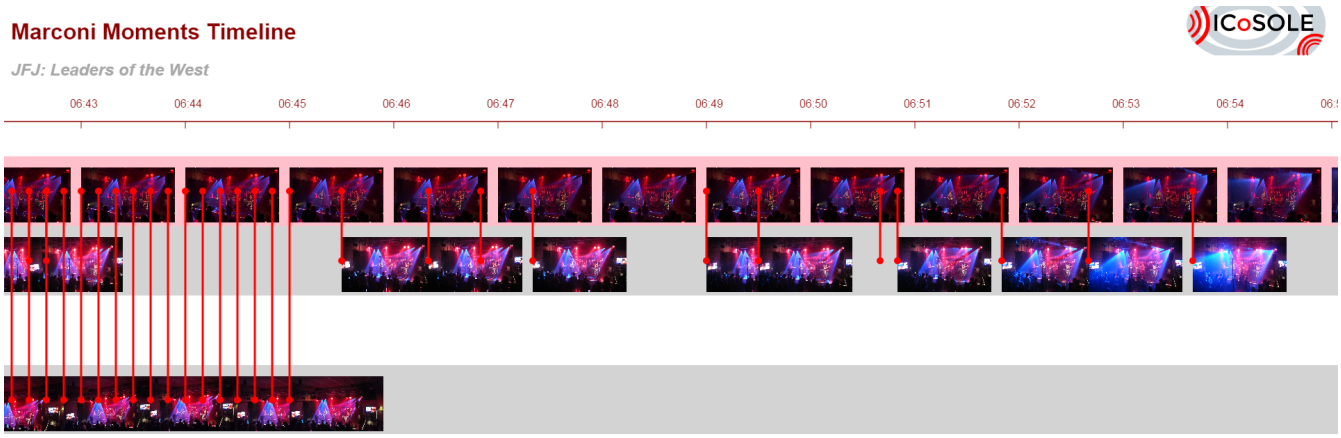

Figure 1. Part of the timeline showing matching results, reference video on top.

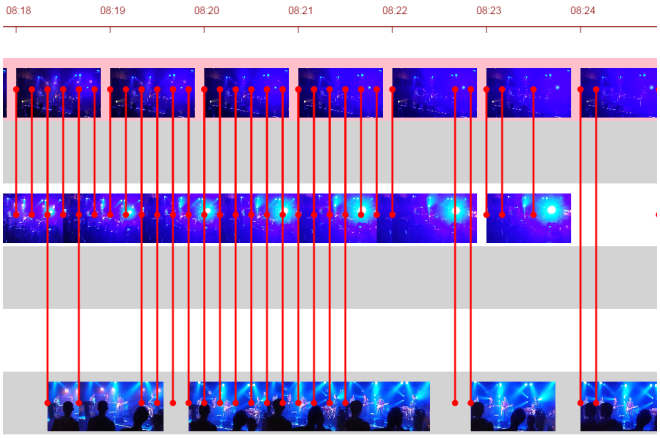

Figure 2. Part of a timeline with matching results.

streams have not been edited, i.e., time is linear throughout the video.

RESULT VISUALISATION

We have evaluated this approach on a data set from an event called Marconi Moments[2], a sequence of two concerts taking place at the radio studio of the Flemish public broadcaster VRT. For synchronization several UGC videos captured within the first nine minutes of the concert were selected in order to temporally align them to one of the professional video streams, in particular, one of the four broadcast cameras providing an overview of the stage. The UGC was recorded by various mobile devices including static and moving sequences, showing the stage and in some cases the audience. The used UGC set contained 9 videos with different lengths, captured at different times in the first minutes of the concert.

We have implemented a web-based visualisation of the matching results, which uses HTML5 canvas and thus only requires only a modern web browser. Each video is shown as a horizontal timeline of key frames. As not all devices are recording all the time, the timelines are not completely filled. The exception is the top-most timeline showing the broadcast reference stream. Matching frames are visualised as red vertical lines, with dots indicating a match in the respective video. One example of the visualisation is shown in Figure 1. Note that matches are not regular over time, as some frames cannot be matched due to strong motion or lights passing through. However, there are more than enough matches to ensure reliable synchronisation and determining the overlapping view. As apparent from the example in Figure 2, the method is reliable enough to work with quite different viewpoints and people occluding part of the stage.

[2] `http://icosole.eu/marconi-moments-test-shoot/`

CONCLUSION AND FUTURE WORK

We have proposed a visual matching approach in order to establish or improve temporal and spatial synchronisation of heterogeneous multi-view video content. We have implemented a web-based viewer to explore the matching results. Future work will address the scalability of the approach (e.g., by using compact visual features in a pre-selection step) in order to enable live application, and integrating the matching results with a live editing application to support content selection.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610370, ICoSOLE ("Immersive Coverage of Spatially Outspread Live Events", `http://www.icosole.eu`).

REFERENCES

1. Fassold, H., and Rosner, J. A real-time GPU implementation of the SIFT algorithm for large-scale video analysis tasks. In *Proc. Real-time Image and Video Processing* (2015).

2. Fürntratt, H., Rosner, J., Stiegler, H., and Fassold, H. GPU-Accelerated SIFT Descriptor Matching. In *GPU Technology Conference* (Mar. 2013).

3. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov. 2004), 91–110.