JRS at Synchronization of Multi-user Event Media Task

Hannes Fassold, Harald Stiegler, Felix Lee, Werner Bailer
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17, 8010 Graz, Austria
{firstname.lastname}@joanneum.at

ABSTRACT

The event synchronisation task addresses the problem of aligning media (i.e., photo and video) streams (“galleries”) from different users temporally and identifying coherent events in the streams. Our approach uses the visual similarity of image/key frame pairs based on full matching of SIFT descriptors with geometric verification. Based on the visual similarity and the given time information, a probabilistic algorithm is employed, where in each run a hypothesis is calculated for the set of time offsets with respect to the reference gallery. From the gathered hypotheses, the final set of time offsets is calculated as the medoid of all hypotheses.

1. INTRODUCTION

The event synchronisation task addresses the problem of aligning media streams (referred to as galleries) from different users temporally and identifying coherent events in the streams. This paper describes the work done by the JRS team for the two subtasks of determining the time offsets of galleries and clustering the images and videos into events. Details on the task and the data set can be found in [1].

2. APPROACH

2.1 Determining Gallery Offsets

Our approach utilizes the visual information (the captured images and extracted key frames from the video) and the given time stamps in a probabilistic way. The absolute time stamps are not considered reliable in this task, however, their relative distances within the gallery of one user can be exploited.

We denote galleries as $\mathcal{G}_{0..M}$ (assuming $\mathcal{G}_0$ as the reference gallery), each $\mathcal{G}_k$ containing a set of images or key frames $I_{1..N_k}$. For every image, several thousands of SIFT descriptors [3] are extracted. A GPU accelerated implementation is used to speed up descriptor extraction and matching [2].

For a pair of galleries (k, l), for each image $I_i \in \mathcal{G}_k$ its best matching image $I_j \in \mathcal{G}_l$ is identified, via exhaustive matching of their respective SIFT descriptors. For each match (I_i, I_j), a geometric verification step is applied, yielding a variable number of homographies along with the number of points h_t supporting the respective homography. The visual similarity $s_{i,j}$ for the image pair is calculated as follows. First, all homographies with $h_t < \tau$ are discarded. From the remaining ones, the k highest values h_t are selected. The selected homographies are clipped to a range $[h_t^{min}, h_t^{max}]$ and the arithmetic average h^{avg} and sum h^{sum} of the clipped values is calculated. The visual similarity $s_{i,j}$ is retrieved as the geometric average of h^{avg} and h^{sum}.

Our general approach is a probabilistic method, where a significant number of potential solutions (hypotheses) are calculated, and from these hypotheses the 'most-inner' (in a sense which will be explained later) is taken as the final solution. Such a probabilistic approach is more robust against outliers in the data. As a preprocessing step, we calculate a *connection magnitude* $c_{k,l}$ for each gallery pair k and l in order to steer the random picking of gallery pairs (k, l) towards the more 'stable' gallery pairs (e.g., the gallery pairs with a high number of matches and a low deviation of the time difference values between the matches). The connection magnitude is calculated as the geometric average of the number of identified matches (based on visual similarity) between the galleries, the average visual similarity scores of the matches and of the reciprocal of the average deviation of the time differences between the matches.

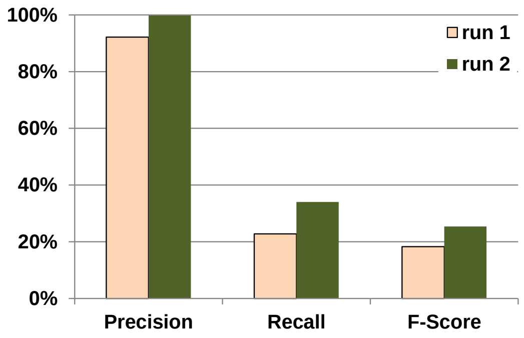
One potential solution is a vector of time differences $D' = (\delta_1, ..., \delta_M)$ between the M galleries and the reference gallery $\mathcal{G}_0$. For generating one potential solution D', we proceed as follows. First, a random gallery pair (k, l) is identified. The probability of picking a pair (k, l) is proportional to its connection magnitude $c_{k,l}$, therefore we steer the random picking towards more stable gallery pairs. In order to probabilistically calculate the time difference $\delta_{k,l}$ between the two galleries, we first apply k-means clustering on the time difference values of all matches, where k is typically in the range 3 to 5. Then, we randomly pick one of the cluster centers and set it as $\delta_{k,l}$. Having calculated $\delta_{k,l}$, we can propagate this value recursively and calculate unknown values $\delta_{k',l}$ by taking usage of the relation

$$\delta_{k,l} = \delta_{k,k'} + \delta_{k',l}, \tag{1}$$

which is very easy to show. By iterating this process of randomly selecting a gallery pair, followed by calculating $\delta_{k,l}$, $M - 1$ times we retrieve one potential solution D'.

In order to calculate the final solution D, we generate a set of several thousands of potential solutions D' (each being a vector of time differences) in the way described above. From the potential solutions, we determine the final solution D by calculating the medoid of all potential solutions. In a certain sense, this is the 'most-inner' solution, when interpreting the potential solutions as vectors.

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany



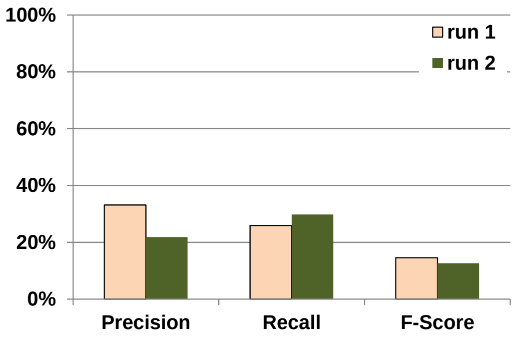


Figure 1: Results for subevent clustering TDF14 (left) and NAMM15 (right).

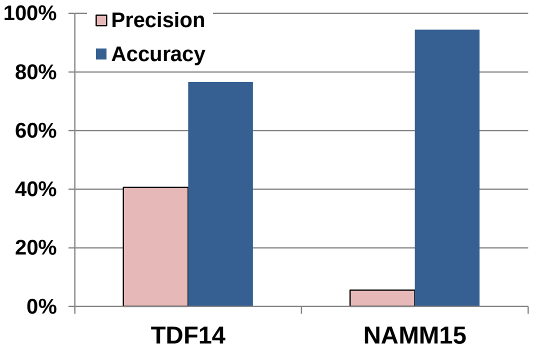


Figure 2: Results for sychronisation.

2.2 Clustering Events

For the event clustering, we rely solely on the time information. We correct the time stamp of a specific gallery with the calculated offset, with respect to the reference gallery, for the specific gallery. Based on the time information, a one dimensional k-means clustering algorithm is applied, where k is ranging between 30 and 100. The value is determined based on the size of a data set - the total number of images in all galleries - and a user parameter which specifies the desired granularity of the subevents.

3. EXPERIMENTS AND RESULTS

We submitted two runs, which use the same parameters for determining the time offsets. The clustering is different, with k for run 2 having the double value of run 1. So run 2 corresponds to a finer granularity of the subevents compared to run 1. Unfortunately, the official submissions only contained the results for the still image data sets (Tour de France, NAMM), but not for the videos.

Figure 2 shows the results for synchronisation. For both data sets, accuracy is clearly higher than precision. This means that our approach tends to optimise for a globally lower synchronisation error at the cost of higher individual errors for some galleries. While precision is significantly lower for NAMM than for TDF, accuracy has actually increased. One reason for this may be the fact, that local features of *bikes* and *bikers* match quite well across many images (which can be seen from the high visual similarity values $s_{i,j}$ for these images), thus visual matching provides a weaker constraint than on visually more diverse data.

The results for subevent clustering are shown in Figure 1. One interesting observation is that while the F1 score is on a comparable level for both data sets, precision and recall are quite balanced for NAMM, but biased towards higher precision for TDF. Interestingly, the variation of the parameter between the two runs does not change this behaviour. For both parameterisations the method tends to oversegment the TDF data. It seems that the impact of synchronisation errors on the clustering result is limited, as no direct relation is apparent from the results.

4. CONCLUSION

The proposed method performs quite well in minimising the overall synchronisation error, but at the expense of more galleries that exceed the error threshold. For the subevent clustering, a better automatic adaptation of the number of clusters to the data set is needed, in order to avoid oversegmentation such as on the TDF data.

Acknowledgments

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/).

5. REFERENCES

[1] Nicola Conci, Francesco De Natale, Vasileios Mezaris, and Mike Matton. Synchronization of Multi-User Event Media at MediaEval 2015: Task Description, Datasets, and Evaluation. In *MediaEval 2015 Workshop*, Wurzen, Germany, September 14-15 2015.

[2] Hannes Fassold and Jakub Rosner. A real-time GPU implementation of the SIFT algorithm for large-scale video analysis tasks. In *Real-Time Image and Video Processing*, San Francisco, CA, USA, 2015.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.