JRS at Event Synchronization Task

Paweł Nowak, Marcus Thaler, Harald Stiegler, Werner Bailer
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17, 8010 Graz, Austria
werner.bailer@joanneum.at

ABSTRACT

The event synchronisation task addresses the problem of aligning photo streams from different users temporally and identifying coherent events in the streams. In our approach, we first determine the visual similarity of image pairs. We determine visual similarity based on full matching of SIFT descriptors and based on VLAD, and compare the use of the two sets of similarity scores. We then build a non-homogeneous linear equation system constraining the time offsets between the galleries based on these matching pairs and determine an approximate solution. Event clusters are initialised from subsequent and visually similar images, and clusters are merged if their temporal proximity and the maximum similarity of their members is high enough.

1. INTRODUCTION

The event synchronisation task addresses the problem of aligning photo streams 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 into events. Details on the task and the data set can be found in [1].

2. APPROACH

2.1 Determining Gallery Offsets

In our approach, we first determine the visual similarity of image pairs. We determine visual similarity based on full matching of SIFT [3] descriptors and based on VLAD [2], and compare the use of the two sets of similarity scores.

The computation of the image similarities between the images of each gallery is based on SIFT descriptors. All images of each gallery were first downscaled from HD to SD. Subsequently, up to 500 SIFT key points and descriptors were extracted from each image.

For similarity calculation based on nearest neighbor matching of SIFT descriptors, each raw SIFT descriptor of the source image is assigned to its nearest neighbour (based on Euclidean distance) descriptor in the target image. These assignments are validated by a homography extracted with the maximum number of descriptors supporting a consistent homography.

For the extraction of the image similarities based on the compact feature representation VLAD the same extracted SIFT key descriptors were used. In order to compute the VLAD signature of each gallery image we used the VLFeat [1] open source library. We reduced a global visual vocabulary with about 300,000 descriptor cluster using k-means clustering to 256 visual words. The descriptors for building the vocabulary have been extracted from a news data set of the TOSCA-MP project[2]. Based on sum of squared errors the similarities between the VLAD signatures and thus the image similarities within the test sets were calculated.

For a pair of images (I_i, I_j), VLAD yields distances d^V_{ij}, which are transformed into similarities

$$s^V_{ij} = \begin{cases} \theta_V - d^V_{ij}, \text{if } d^V_{ij} < \theta_V \\ 0, otherwise, \end{cases} \tag{1}$$

where θ_V is a threshold for the maximum distance. The SIFT similarity s^S_{ij} is determined as

$$s^S_{ij} = \begin{cases} \max(0, \frac{|P_{ij}|}{\min(|P_i|,|P_j|)} - \theta_S), \text{if } |P_{ij}| \geq p \\ 0, otherwise, \end{cases} \tag{2}$$

where P_i are the key points in each of the images, P_{ij} is the set of matching key points, p is a threshold for the number of matching key points and θ_S is a similarity threshold. We use all similarities above zero to formulate constraints on the time offsets of the galleries. Optionally, the GPS information of the images (if available) can be used, setting the similarity to zero, if the deviation in longitude or latitude is above a threshold θ_G (in degrees).

For N galleries $G_1, \ldots, G_N$, we can assume without loss of generality that G_1 is the reference gallery. We aim at obtaining a list of time differences $D = (\delta_2, \ldots, \delta_N)$, where δ_i is the time offset between galleries G_i and G_1. As the underlying assumption in this task is that the offset between two galleries is constant over time, each pair of matching images adds one constraint of the form $\delta_p - \delta_q = \tau_{ij}$, where p, q are the galleries containing images I_i, I_j respectively, and τ_{ij} is the time offset determined from time stamps of the matching images. Note that δ_1 is by definition 0. We can then reorganise our constraints into an overdetermined equation system

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

[1] http://www.vlfeat.org
[2] http://www.tosca-mp.eu

run	set	vis.sim.	θ_S	θ_V	θ_G	α_t
1	1,2	VLAD	n/a	1.70	2.5	1.0
2	1	SIFT+VLAD	0.07	1.82	2.5	0.0
2	2	SIFT+VLAD	0.08	1.80	2.5	1.0
3	1	SIFT+VLAD	0.07	1.82	2.5	0.0
3	2	SIFT+VLAD	0.08	1.85	2.5	0.0
4	1	SIFT+VLAD	0.06	1.85	2.5	0.0
4	2	SIFT+VLAD	0.08	1.80	2.5	0.0

Table 1: Parameters of runs, $t_{min} = 120s$ and $p = 10$.

$$\begin{bmatrix} g_2(i) - g_2(j) & \cdots & g_N(i) - g_N(j) \\ \vdots & & \vdots \\ g_2(k) - g_2(l) & \cdots & g_N(k) - g_N(l) \end{bmatrix} \begin{bmatrix} \delta_2 \\ \vdots \\ \delta_N \end{bmatrix} = \begin{bmatrix} \tau_{ij} \\ \vdots \\ \tau_{kl} \end{bmatrix} \tag{3}$$

where $g_n(i)$ is a binary function, yielding 1 if $I \in G_n$, 0 otherwise. In order to deal with outliers, we iteratively solve the equation system, and remove up to 10% of the largest outliers. In each iteration, we use the Jacobi method to solve the equation system.

2.2 Clustering Events

We initialise the event time line by grouping subsequent images, which have visual similarity (s^V or s^S) above zero. This will oversegment the event time line. In a next step, we start regrouping these events based on visual similarity and (optionally) temporal proximity. The distance between two events i, j is determined as

$$d^E_{ij} = \alpha_t \text{max}(1, \frac{|\bar{t}_i - \bar{t}_j|}{t_{min}})\theta - \text{max}_{k \in E_i, l \in E_j} s_k l, \tag{4}$$

where $\bar{t}_i$ is the mean time of images in event E_i, α_t is a weight for using time information, θ is the similarity threshold used (S or V) and $s_k l$ is the visual similarity between a pair of images of which one belongs to E_i and the other to E_j. Two events are merged if $d^E_{ij} < \theta_{merge}$, where θ_{merge} has been set to $\theta_V + 0.15$.

3. EXPERIMENTS AND RESULTS

We submitted four runs, with the parameters listed in Table 1. One observation of the experiments of the test set is that full matching of SIFT descriptors is better for determining gallery offsets, which needs to find the single most similar image from the other gallery. In contrast, the event clustering needs a more global notion of similarity, which is well covered by VLAD. Thus we used VLAD similarities for event clustering in all the runs. The results for synchronisation are shown in Figure 1, and those for clustering in Figure 2.

4. DISCUSSION

As already expected from the experiments on the development set, VLAD is not discriminative enough for determining the image pairs for synchronisation, thus the results of run 1 are much worse than the others. Our method only manages to sychronise a fraction of the galleries correctly, however, if a gallery is sychronised, the accuracy is rather high. The results for the London data set are clearly better than those for the Vancouver set. We think that this is not so much related with the similarity to the development set, but rather with the high visual similarity in Winter Olympics (e.g., all ice based competitions have high similarity). For clustering, the differences are not so clear, for runs 2 and 3 the Vancouver results are even better than the London ones according to Jaccard index and F-measure. In general, the Rand index shows a quite different picture than the other two measures.

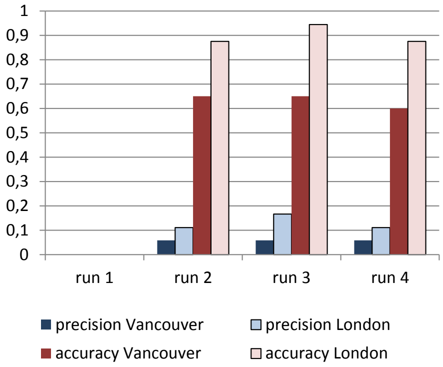

Figure 1: Results for sychronisation.

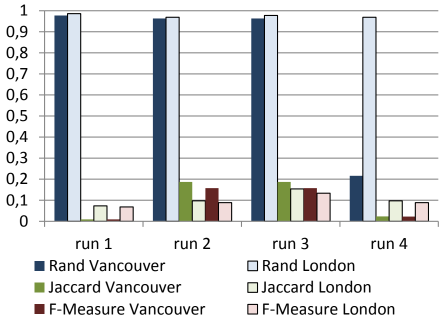

Figure 2: Results for clustering.

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, and Vasileios Mezaris. Synchronization of Multi-User Event Media (SEM) at MediaEval 2014: Task Description, Datasets, and Evaluation. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[2] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

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