Estimation of correspondent trajectories in multiple overlapping synchronized videos using correlation of activity functions

Thierry Malon, Sylvie Chambon, Vincent Charvillat, Alain Crouzil

ESTIMATION OF CORRESPONDENT TRAJECTORIES IN MULTIPLE OVERLAPPING SYNCHRONIZED VIDEOS USING CORRELATION OF ACTIVITY FUNCTIONS

Thierry Malon, Sylvie Chambon, Vincent Charvillat, Alain Crouzil

IRIT, Université de Toulouse, Toulouse, France

ABSTRACT

We present an approach for ranking a collection of videos with overlapping fields of view. The ranking depends on how they allow to visualize as best as possible, i.e. with significant details, a trajectory query drawn in one of the videos. The proposed approach decomposes each video into cells and aims at estimating a correspondence map between cells from different videos using the linear correlation between their functions of activity. These latter are obtained during a training session by detecting objects in the videos and computing the coverage rate between the objects and the cells over time. The main idea is that two areas from two different videos that systematically offer presence of objects simultaneously are very likely to correspond to each other. Then, we use the correspondence between cells to find the reformulated trajectory in the other videos. Finally, we rank the videos based on the visibility they offer. We show promising results by testing three aspects: the correspondence maps, the reformulation and the ranking.

***Index Terms*—** Trajectory reformulation, video surveillance, multiple views, overlapping fields of view, matching

1. INTRODUCTION

The multiplication of multimedia devices allowing to record videos raises new challenges. Nowadays, it is easy to find multiple videos from the same event. Multiple views with overlaps offer a richer understanding of the scene compared to single view. However, manually watching each video is a long and tedious task. Consequently, there is a need in helping users to easily navigate through a collection of videos.

In recent years, numerous works proposed approaches to tackle the challenge of easing multiple video visualization. When camera calibration parameters are known and numerous images of the scene are available, a static 3D reconstruction of the scene can be obtained by detecting and matching keypoints [1]. To incorporate the temporal aspect of synchronized videos, [2] proposed a 4D reconstruction of the scene with both the static parts and also the dynamic parts that are moving over time. These reconstructions capture elements from multiple viewpoints and thus provide a good overall representation and understanding of the scene. However, they cannot always be performed as they require camera calibration parameters and low viewpoint variations between the cameras. When 3D reconstruction is not feasible, navigation through videos by switching over time to the one that best describes the scene was investigated. In [3], scores are attributed to each view using the activity of the objects, their size, location and number, as well as particular events occurring. The higher the score, the most interesting the view. In this paper, we suppose that calibration, and so 3D reconstruction, is not possible, due to uncontrolled acquisition of the videos.

In video surveillance, investigators generally have hundreds of cameras to process, some of which can present overlaps in their fields of view. To navigate through the videos, they want to be able to manually formulate a query on elements of interest and choose among a ranked list of videos that match the query. When dealing with overlapping videos, in [4], the user can query a region of interest of the current video and be redirected to the video that offers the best entropy regarding the objects contained in the queried region.

For cases where the camera views are disjoint, the authors of [5] proposed to jointly compute the camera network topology and the re-identification. They iteratively refine topology by using re-identification and reversely. In the same context, [6, 7] proposed to study the structure of the camera network by estimating links between the disjoint cameras and the time delay between them. They cut each view into cells, measure the activity over time of each cell using a training set made of videos with detections and then merge cells with similar activity into regions. Their Cross Canonical Correlation Analysis applied to functions of activity showed good results for estimating the spatial and temporal topology inference of a camera network. Consequently, we propose to estimate correspondences between multiple videos presenting overlaps by using functions of activity inspired by this measure.

Most approaches based on multiple cameras with overlapping fields of view suppose that it is possible to estimate the camera calibration parameters. In general cases, i.e. with videos from surveillance cameras combined with smartphone videos, it becomes difficult to have a reliable estimation of these parameters. In this work, we do not compute calibration parameters. As far as we are concerned, in the same context, we can cite [8] where the lines of view of cameras presenting overlapping fields of view are retrieved from a training set of videos by detecting the point on the ground plane of a

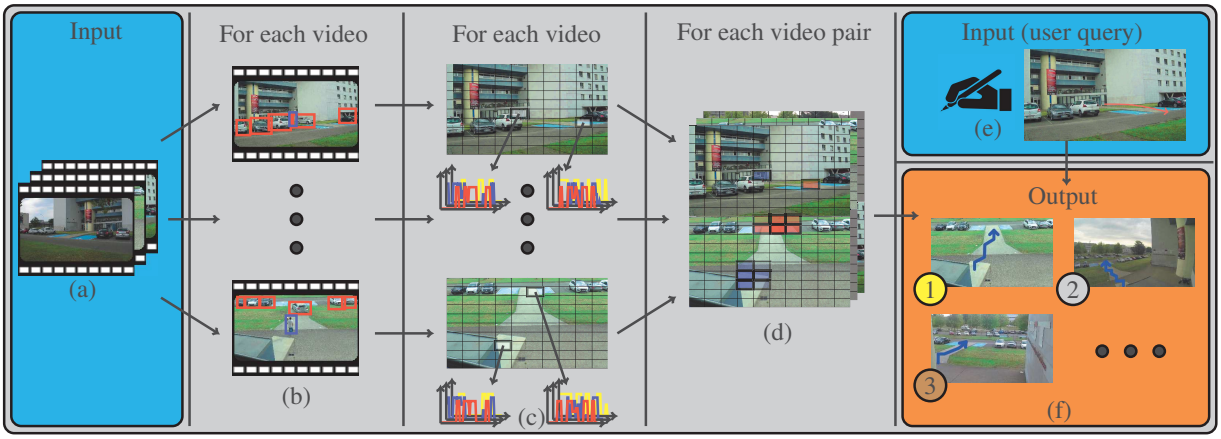

Fig. 1: General overview of the approach: (a) Collection of videos as input, (b) detection of objects and categories, (c) functions of activity computing, (d) correspondence maps computing, (e) user trajectory query, (f) video ranking based on visibility score.

detected object in a view at the moment when it appears in another view. In this paper, we also use a training phase.

The main contribution of this paper is to propose a new approach for reformulating a trajectory query drawn in one view into its corresponding trajectories in other overlapping views, and then for ranking these views regarding the visibility of the scene they offer. We first introduce functions of activity for estimating correspondence maps between video pairs. Then, when a user draws a query trajectory in one video, we propose a reformulation score to find the corresponding trajectories in the other overlapping videos. Finally, we define a visibility score to rank these matched videos regarding how spread the trajectories are in each video as we assume that the longer the trajectory, the more interesting the view, i.e. the view shows more details about the trajectory than the initial view.

2. PROPOSED METHODOLOGY

We suppose that the videos are shot from static viewpoints and are temporally synchronized, i.e. that they start at the same timestamp and have the same duration. We have no calibration parameters. The approach consists of the following steps, see Figure 1. For each video, see (a), we first detect the objects of interest and assign them a category, see (b) where red bounding boxes belong to vehicle and the blue one to a pedestrian, which can be done using ROLO [9]. Then, in (c), we decompose each view into cells (the grid on the figure), and associate to each cell a function of activity per object category defined as its occupation rate over time. Then, for each video pair, in (d), we assign to each cell of the first video a correspondence map regarding the correlation of its occupation rate over time (two particular cells are highlighted in the top view as well as their corresponding cells in the bottom view). Our main idea is that, if two cells from two different synchronized cameras are systematically simultaneously occupied, they are likely to correspond to the same 3D position. Thus, when drawing a query trajectory in a view, in (e) (the red bold line), we list the cells crossed by this query and look for the corresponding cells in the other view. For each other view, the reformulated trajectory is obtained using the sequence of cells that has the best reformulation score, defined as a trade-off between maximizing correspondences and minimizing the distances between consecutive cells. A ranking of the videos about the visibility of the trajectory is finally returned in (f) based on the reformulation score and the visibility of the reformulated trajectory.

2.1. Correspondence maps

In this section, we define the notion of function of activity and how we compute the correspondence map from a region in one videos to its corresponding regions in another video. Let V_1 and V_2 be two temporally synchronized camera views with overlapping fields of view. We note $d_\omega^V(t)$ the ensemble of detections containing all the objects of category ω that are detected at time t in video V, used to learn the correspondence maps. As clusters of close pixels generally correspond to the same region, we divide each view V into N cells $c_\mathbf{i}^V$ of identical size. The impact of N is measured in Section 3.

The function of activity of category ω in a view V, denoted $a_\mathbf{i}^{V,\omega}$, is the recovery rate of the cell by the detections of category ω objects d_ω^V over time. Note that each cell has one function of activity per category of object (see Figure 1.(c)). We define the correspondence rate of category ω between two cells of two different videos as:

$$\mathcal{C}_\omega(c_\mathbf{i}^{V_1,\omega}, c_{\mathbf{i}'}^{V_2,\omega}) = \max(0, \text{corr}(a_\mathbf{i}^{V_1,\omega}, a_{\mathbf{i}'}^{V_2,\omega})) \quad (1)$$

where corr stands for the linear correlation operator between

two random variables. For each cell $c_{\mathbf{i}}^{V_1,\omega}$, we also define a correspondence map $\mathcal{C}_\omega(c_{\mathbf{i}}^{V_1,\omega}, V_2)$ containing the correspondences between $c_{\mathbf{i}}^{V_1,\omega}$ and all the cells of V_2. Figure 1.(d) illustrates the correlation maps: a pair of views is depicted, and two cells are highlighted in blue and red (respectively for categories "person" and "car") in the top video. The bottom view shows another video with the correspondence maps of the two cells. The main idea for matching cells is that corresponding cells are expected to present correlated functions of activity. In fact, it is unlikely that a pair of non corresponding cells systematically presents the same activity over time.

2.2. Trajectory reformulation

We propose a trajectory reformulation scheme relying on the cell correspondences. A trajectory is defined as a succession of connected 2D segments.We extract the sequence $\mathcal{S}$ of M cells $(c_{\mathbf{i}_1}, ..., c_{\mathbf{i}_M})$ that are crossed by the segments with no consecutive identical cells.

To find the corresponding trajectory, we want to find the successive indices of the cells in the other view $(\mathbf{i'}_1, ..., \mathbf{i'}_M)$ that maximize the sum of the correlations with the successive crossed cells while ensuring a continuity in the reformulated trajectory. To this end, we penalize successive corresponding cells that are not adjacent regarding their distance and we define the reformulation score of the sequence of cells $\mathcal{S}$ between views V_1 and V_2 as:

$$\underset{(\mathbf{i'}_1,...,\mathbf{i'}_M)}{\operatorname{argmax}} \frac{\frac{1}{M}\sum_{k=1}^{M} \mathcal{C}_\omega(c_{\mathbf{i}_k}^{V_1}, c_{\mathbf{i'}_k}^{V_2})}{1 + \sum_{k=1}^{M-1} \max(0, ||\mathbf{i'}_k - \mathbf{i'}_{k+1}|| - 1)} \quad (2)$$

The numerator favors cells of V_2 that have a good activity correlation with the cells crossed by the query in V_1 while the denominator penalizes consecutive cells that are not adjacent in the reformulated trajectory. We obtain the reformulated trajectory in the other view by joining the centers of the sequence of cells of index $(\mathbf{i'}_1, ..., \mathbf{i'}_M)$.

2.3. Selection and ranking of videos

In this section, we explain the proposed method for selecting the videos that are related to the queried view and ranking them regarding the visibility of the reformulated trajectory they offer. As stated before, investigators often have to treat dozens of videos at once. When focusing on a particular location, they may want to automatically navigate between the videos of the same scene, i.e. views presenting overlaps. It is very unlikely that a view with no overlap with the queried view contains any region presenting correlated activity with a region of the queried view for the whole duration. Thus, views where the trajectory query cannot be reformulated suffer from a low reformulation score and can be filtered using a threshold σ. The remaining reformulated trajectories are then ranked regarding their visibility score that we define as the product of their total length and their reformulation score. We choose these criteria because we suppose that the longer the trajectory, the higher the number of details that can be seen.

3. EXPERIMENTS

We want to evaluate the quality of the correspondence maps, the trajectory reformulation and the ranking of the views that offer a better visualization. We used the ToCaDa dataset [10]. It contains 25 videos in which about 30 objects of 3 categories (person, motorcycle and car) are present. Among all the videos, 15 views present large overlaps while the others are shooting non overlapping areas. The videos are temporally synchronized and have the same duration (≈5 minutes).

We first evaluate the quality of the correspondence maps between the 15 videos with overlaps by measuring how much the objects that cover cells of a video also covers the corresponding learned cells in the other videos. For each pair of videos (V_1, V_2), we list the objects that simultaneously appear in both videos and compute all the bounding boxes at times of simultaneous presence. Then, we define the correspondence rate between each pair B_1 and B_2 of simultaneous bounding boxes of a same object of category ω seen in V_1 and V_2:

$$\sum_{(\mathbf{c_i}^{V_1,\omega}, \mathbf{c_j}^{V_2,\omega})} \frac{|c_{\mathbf{i}}^{V_1,\omega} \cap B_1|}{|B_1|} \frac{|c_{\mathbf{j}}^{V_2,\omega} \cap B_2|}{|B_2|} \frac{\mathcal{C}_\omega(c_{\mathbf{i}}^{V_1,\omega}, c_{\mathbf{j}}^{V_2,\omega})}{||\mathcal{C}(c_{\mathbf{i}}^{V_1,\omega}, V_2)||} \quad (3)$$

The first term evaluates how the bounding box in the first camera is covered by the cell $c_{\mathbf{i}}^{V_1,\omega}$ whereas the second term evaluates the same aspect for the corresponding cell $c_{\mathbf{j}}^{V_2,\omega}$ with the corresponding bounding box in the second camera. The last term is related to equation (1). The mean correspondence rate is obtained by averaging over all simultaneous bounding boxes of a same object, over all objects in a pair of videos and over every pair of videos. We used the provided category labels, detection and tracking of the bounding boxes of the objects in each view to compute the correspondence maps of the cells. Figure 3 presents the correspondence rates for different setups and different number of cells. We tried to evaluate the behavior of the method when categories are not distinguished, when adding a temporal desynchronization of one second between each pair of videos or when training is done on only half of the dataset. As expected, the results reveal that this step is quite sensible to perturbations. Also, distinguishing between the categories widely improves the correspondence map rate. Note that we do not expect high correspondence rates as the correspondence map of one cell in a video generally covers a large amount of cells in the other videos. However, we expect that this correspondence rate is sufficient for the next step evaluated: the quality of the reformulation.

Fig. 2: Best views ranking. 1st column: three trajectory queries are drawn in red, respectively for categories human, moto and car. 2nd to 4th columns: the top 3 views that offer both a high visibility score and where the reformulated trajectory occupies most space are returned in descending order. 5th column: an overlapping view with a low rank.

We drew 10 trajectory queries at the ground level in different overlapping views and applied the proposed reformulation method. To estimate a corresponding trajectory of reference, i.e. to be compared with, we computed the homographies between the ground plane of each pair of overlapping views by using the corners of the blue parking space. Then, we measured the Dynamic Time Warping (DTW) distance [11], in pixels, between the two trajectories (obtained by applying the homography and using our method). Figure 4 presents the mean DTW distance on videos of size 960×540 for different numbers of cells. Again, as expected, the reformulation becomes reliable when the number of cells raises and these results validate the quality of the proposed reformulation.

For the best view ranking, the views that present no overlap with other videos are correctly filtered when using $\sigma = 0.3$ as almost no correspondence can be learned due to the absence of systematic simultaneous presence of objects. Figure 2 presents the top 3 best views proposed for different trajectory queries drawn in different views. An overlapping view with a low visibility score is also presented and mostly corresponds to views where the trajectory is not fully visible or short due to the viewpoint of the camera. Among the top 5 views returned on our 10 trajectory queries, 72% of the views give as much as or more visibility to the trajectory.

4. CONCLUSION

From a collection of videos with no calibration, the proposed method allows to successfully rank a subset of videos that present overlaps in their fields of view regarding the visibility they offer of a trajectory query. Future work may deepen this method by relaxing the constraint of synchronization and estimating the time delay between different videos, using photometric information of the objects and exploiting the neighborhood of the cells to compute the correlation maps.

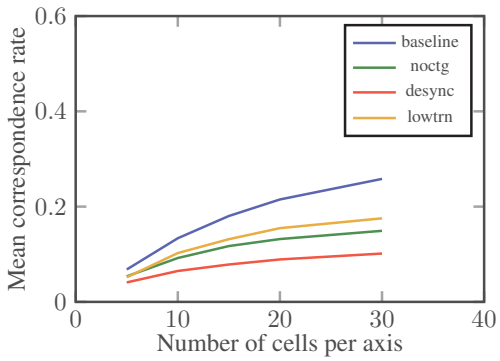

Fig. 3: Mean correspondence rate on the ToCaDa dataset [10] for different number of cells and different setups: baseline, without distinguishing the categories (noctg), with a 1 second desynchronization between the videos (desync) and training on only half the dataset (lowtrn).

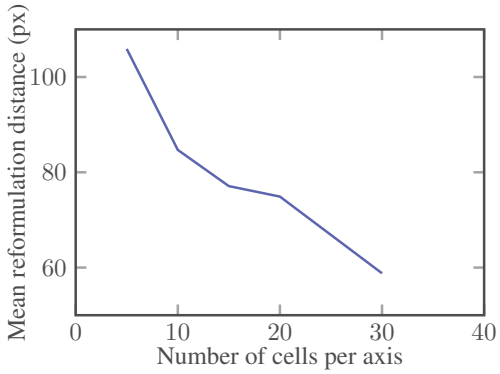

Fig. 4: Mean reformulation DTW distance in pixels for different number of cells.

5. REFERENCES

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *IEEE International Conference on Computer Vision*, 2009.

[2] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton, "Temporally coherent 4D reconstruction of complex dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[3] F. Daniyal, M. Taj, and A. Cavallaro, "Content and task-based view selection from multiple video streams," *Multimedia tools and applications*, 2010.

[4] A. Carlier, L. Calvet, P. Gurdjos, V. Charvillat, and W. T. Ooi, "Querying multiple simultaneous video streams with 3D interest maps," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. 2017.

[5] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon, "Joint person re-identification and camera network topology inference in multiple cameras," *Computer Vision and Image Understanding*, 2019.

[6] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *International Journal of Computer Vision*, 2010.

[7] C. C. Loy, T. Xiang, S. Gong, et al., "Incremental activity modeling in multiple disjoint cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[8] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[9] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking," in *IEEE International Symposium on Circuits and Systems*, 2017.

[10] T. Malon, G. Roman-Jimenez, P. Guyot, S. Chambon, V. Charvillat, A. Crouzil, A. Péninou, J. Pinquier, F. Sèdes, and C. Sénac, "Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views," in *ACM Multimedia Systems Conference*, 2018.

[11] Z. Zhang, K. Huang, and T. Tan, "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes," in *IEEE International Conference on Pattern Recognition*, 2006.